Peer feedback provision and mathematical proofs: Role of domain knowledge, beliefs, perceptions, epistemic emotions, and peer feedback content

Maryam Alqassab

München 2016

Peer feedback provision and mathematical proofs: Role of domain knowledge, beliefs, perceptions, epistemic emotions, and peer feedback content

Maryam Alqassab

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

am Munich Center of the Learning Sciences

der Ludwig-Maximilians- Universität

München

Vorgelegt von

Maryam Alqassab

München, den 10. November 2016

1st Supervisor / Erstgutachter: Prof. Dr. Jan-Willem Strijbos

2nd Supervisor / Zweitgutachter: Prof. Dr. Stefan Ufer

Tag der mündlichen Prüfung: 03.02.2017

The research presented in this work was supported by the Elite Network of Bavaria [Project number: K-GS-2012-209]. I would like to extend my sincere gratitude to the opportunities (conference attendance, incubator stay, international knowledge exchange) made possible by the ENB.

Acknowledgements

This dissertation signals the end of plenteous three years that has been a period of vigorous growth for me, not only scientifically but also as a person. This accomplishment would not have been possible without the valuable contributions of a number of people that I would like to acknowledge here.

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Dr. Jan-Willem Strijbos and Prof. Dr. Stefan Ufer. Your endless support and guidance helped me throughout the hard times of conducting my research and writing this dissertation. Due to your stimulating feedback, I became more aware of the strengths and weaknesses of my work. It has been a very enriching experience to work with two researchers from different domains. I could not have asked for a better supervisory team!

Jan-Willem, you have been a tremendous mentor for me. The long hours of discussing the little details of my research, helping with coding, and observing my culture shocks and my Ph.D. blues with a smile on your face have been all priceless. Yet, it was not always easy to keep up with your 'not-do' list; I have to admit! Thanks for giving me the freedom to make ambitious decisions regarding my research, yet stepping in willingly whenever it was needed. And special thanks for pushing me beyond my limits!

Prof. Ufer, it gives me a great pleasure to acknowledge all the help and support you provided me throughout my Ph.D., starting with creating materials for my studies, to providing domain knowledge support, ensuring that data collection at the Mathematics department runs smoothly, and of course providing me the opportunity to finish writing my dissertation in Germany.

My sincere thanks also goes to Prof. Dr. Frank Fischer for different kind of support and resources he provided to ensure that my studies take place on time and to all of his chair members for their feedback and help especially during the absence of Jan-Willem. A special thanks to Markus Bolzer for his valuable help with the eye-tracker.

I would like to thank my REASON colleagues for making the Ph.D. an exciting experience. The group diversity taught me a new lesson every day. My office buddies, Ansgar, Sandra, and Janina thanks for helping me during my struggles with German bureaucracy and for being very polite all the time and not speaking German when I was around. Although you tried to teach me German, the "word of the day" did not last long enough. So I can blame you for not learning German for the rest of my life! Özgün and Maria, thanks for your emotional support, back exercise breaks, and the Maths department lunch gatherings during the last month. Strobel, I guess it's time to join my "imaginary husband in Canada"…you will always remain a unique friend! Adras, I can finally say that I will miss the bad jokes. Katharina, I will remember you every time I watch Sherlok Holmes. Chris Ghanem, you were a perfect example to fight my stereotypes about Germans. Stay as you are! And of course, the great two guys who helped me a lot, Daniel and Mauro. Daniel, I am grateful for your kind help with reviewing the mathematical tasks and your endless support at the Mathematics department. Mauro, thanks for helping me with the R codes even when you were extremely busy. You still have to teach me knitting, maybe via skype!

A special thanks to my international supervisor Prof. John Hattie for his generous support and to his team at the SLRC for the great time, feedback, and motivation I received during my research incubation stay in Melbourne. I consider this experience as a turning point in my project!

I would like to thank all of the student assistants who helped me with my project. Fabian, Alex, Marietta, and Wilken, it would not have been possible to finish my project without your great work.

I am also indebted to my German family the Teucherts, for treating me like their daughter. Your support made it easier for me to settle down in Munich and to live a comfortable life. You are one of the reasons that will make me come back to visit Munich!

Last but not least, I would like to thank my family for always being there for me through my good and bad times no matter how far or close I am. My words cannot express how grateful I am to my dear husband Mojtaba who did not hesitate for a second to endure a long distance marriage so that I can fulfill one of my dreams! To Mojtaba and my family, I dedicate this dissertation.

Maryam

November 2016

Table of Contents

Summary

Peer feedback (PF) is nowadays frequently used as an “assessment for learning” activity in higher education. Despite the growing body of research about PF in different domains, the primary focus of research seems to be on the PF message. Students’ individual characteristics likely to influence the PF activity including domain knowledge, beliefs about PF provision, PF providers’ perceptions of their PF, and epistemic emotions are still widely ignored in PF research. Providing PF is considered beneficial for students, yet studies exploring processes and individual characteristics underlying PF provision are limited. In mathematics education, reasoning and proofs are regarded central to students’ learning (National Council of Teachers of Mathematics, 2000). Preservice mathematics teachers are, thus, expected to acquire assessment skills to equip them to assess students’ mathematical arguments in their future classrooms. PF provision can be a useful activity to train preservice mathematics teachers’ assessment skills of complex geometric tasks that involve scientific reasoning and argumentation. However, research in this area is still limited. This dissertation aims to explore the mechanisms and benefits of PF provision by preservice mathematics teachers on geometric construction tasks and geometry proofs performed by (fictional) peers. More specifically, this dissertation investigates PF provision by exploring: (a) the PF composition process, (b) preservice mathematics teachers’ individual characteristics (domain knowledge, beliefs about PF provision, PF provider’s conceptions of PF, and experienced epistemic emotions), and (c) the relationships between preservice mathematics teachers’ individual characteristics and the content of provided PF. For this purpose, two empirical studies were conducted.

In Study 1, the impact of a structured PF training on preservice mathematics teachers’ PF provision skills and their beliefs about PF provision was investigated, taking into account their domain knowledge. A quasiexperimental field study was conducted with a mixed design.

Forty-three preservice mathematics teachers received PF provision training on geometric construction tasks after measuring their basic PF provision skills, their basic geometric knowledge and their beliefs about PF provision. Based on their performance in the basic geometric knowledge test, the participants were grouped into three domain knowledge groups (low, medium, and high) representing the lowest, middle, and highest one-third of the sample (between-subject factor). All participants received PF provision training – based on Hattie and Timperley (2007) feedback model – over two fifty-minute sessions. In these sessions, the preservice mathematics teachers were introduced to the Hattie and Timperley (2007) feedback model that differentiates between progressive levels of feedback (task-level, process-level, self-regulation level, and self-level), and they received PF provision prompts and an evaluation rubric. The participants worked on several activities to familiarize themselves with the different feedback levels in the Hattie and Timperley (2007) model. The preservice mathematics teachers then were involved in two practice sessions in which they provided written PF on fictional peer solutions to geometric construction tasks with the help of PF provision prompts, an evaluation rubric, and a worked example. Preservice mathematics teachers' PF skills and their beliefs about PF provision were measured again at the end of the semester (within-subject factors).

The results of Study 1 revealed that the structured PF training was beneficial for preservice mathematics teachers, yet their domain knowledge influenced their gains from the training. Only preservice teachers with medium and high domain knowledge provided more self-regulation level PF, which is the highest level of the Hattie and Timperley (2007) model. Conversely, preservice mathematics teachers with low domain knowledge provided more task-level PF than those with medium domain knowledge after the training. No significant changes were observed in process-level and self-level PF after the training for all domain knowledge groups. Additionally, preservice mathematics teachers' beliefs about PF provision became less positive after the PF training regardless of their domain knowledge. It was concluded from Study 1 that a basic level of domain knowledge is required to train higher levels of PF on peer solutions to geometric construction tasks and that preservice mathematics teachers' beliefs about PF provision become more realistic after being involved in PF provision activities.

In Study 2, the PF composition process was explored through monitoring preservice mathematics teachers' eye-movements while they provided PF on a fictional peer solution to a geometry proof. The study investigated the impact of peer solution quality (near-correct vs. erroneous) on the proportional total dwell time (PTDT) on different components of the peer solution (text vs. figure), PF provider's proof comprehension, PF content, and experienced curiosity and confusion. The study also investigated the relationships between PTDT, proof comprehension, PF content, epistemic emotions curiosity, confusion, and anxiety, as well as, beliefs about PF provision and PF providers' perceptions of their PF.

In a between-subject experimental design, 53 preservice mathematics teachers were randomly assigned to a near-correct (fictional) peer solution condition (*n* = 27) or an erroneous (fictional) peer solution condition (*n* = 26). Each participant had to provide verbal PF on the peer solution while their eye-movements were recorded using a head-mounted eye-tracker (on-screen phase). The fictional peer solutions were designed in a way that they had an earlier non-comparable part that is erroneous for one condition and near-correct for the other, and a comparable (later) part that is the same for both conditions. In the second phase (paper-and-pen) the participants answered a proof comprehension test, a basic geometric knowledge test, and they completed epistemic emotions, beliefs about PF provision, and PF providers' perceptions of their PF questionnaires. The PF content was operationalized in this study along two dimensions, PF type (cognitive surface, cognitive verification, cognitive elaboration, self-efficacy, and affective) and PF accuracy.

An effect was found for peer solution quality on PTDT on different components of the peer solution. For the non-comparable peer solution part, preservice mathematics teachers who provided PF on the erroneous peer solution had longer PTDT on the text and the figure of the peer solution compared to those who provided PF on the near-correct peer solution. For the comparable peer solution part, preservice mathematics teachers who provided PF on the erroneous peer solution had longer PTDT on the text of the peer solution, whereas those who provided PF on the near-correct peer solution had longer PTDT on the figure of the peer solution. It was suggested that providing PF on an erroneous peer solution to geometry proof might stimulate adopting an analytical approach, whereas providing PF on a near-correct peer solution to geometry proof might stimulate a figure-based approach. No significant relationships were found between PTDT on different components of the peer solution and PF provider's proof comprehension, but participants who provided PF on the near-correct peer solution had a better proof comprehension than those who provided PF on the erroneous peer solution, even after controlling for their basic geometric knowledge.

No effect of peer solution quality was found on PF types (cognitive surface, cognitive verification, cognitive elaboration, and self-efficacy). Conversely, preservice mathematics teachers in the near-correct peer solution condition provided more accurate PF than those who were in the erroneous peer solution condition, after controlling for their basic geometric knowledge. It was concluded that it is easier for preservice mathematics teachers to validate correct proofs than erroneous proofs, which is consistent with previous findings (e.g., Healy & Hoyles, 1998; Reiss, Heinze, & Klieme, 2000).

For the erroneous peer solution condition, PTDT on the text of the non-comparable peer solution part was positively moderately related to cognitive verification PF. For the near-correct peer solution condition, PTDT on the text of the non-comparable peer solution part was negatively moderately related to self-efficacy PF, while PTDT on the figure corresponding to

the non-comparable peer solution part was positively moderately related to self-efficacy PF. For the same condition, moderate negative correlations were found between PTDT on the text of the comparable peer solution part and cognitive elaboration PF and PF accuracy, and moderate positive correlations were found between PTDT on the figure corresponding to the comparable peer solution part and cognitive elaboration PF and PF accuracy. It was suggested that figure-based approach might facilitate providing more cognitive elaboration PF and more accurate PF, yet this seems to be the case only for the near-correct peer solution. No significant relationships were found between proof comprehension and PF types, but proof comprehension was positively related to PF accuracy for both conditions.

Preservice mathematics teachers in the erroneous peer solution condition experienced more curiosity during PF provision than those in the near-correct peer solution condition, but no significant differences were found in confusion. Confusion and anxiety were negatively related to PF accuracy but only for the near-correct peer solution condition. PF accuracy was positively (moderately to strongly) related to PF providers' perceptions of their PF for the near-correct peer solution condition. Additionally, preservice mathematics teachers' perceptions of their PF were positively (moderately to strongly) related to their confidence regarding PF provision for both conditions. Overall, it was suggested that preservice mathematics teachers' confidence regarding PF, their perceptions of their PF, and epistemic confusion and anxiety are likely to be indicators of the accuracy of their PF.

1. General Introduction

> Creative, exploratory learning requires peers currently puzzled about the same terms or problems. Large universities make the futile attempt to match them by multiplying their courses, and they generally fail since they are bound to curriculum, course structure, and bureaucratic administration. (Ivan Illich, 1971, p. 10)

The complexity of our world makes assessment skill a requirement for students to thrive out of school. Students as lifelong learners should be able to monitor and evaluate their learning and that of others, a skill that is often ignored even in university education (Boud & Falchikov, 2006). In mathematics education, assessment is regarded as a key component to help students become independent learners (National Council of Teachers of Mathematics, 2000), and one way to activate students' role in assessment is through involving them in peer feedback (PF). Students need peers of equal status who live similar learning experiences, to share their insecurities and become more aware of their strengths and weaknesses. PF activities are crucial for teacher-training programs because preservice teachers are required to develop adequate assessment skills to enable them to assess their future students' learning and to provide feedback to their students and their peers. Students' individual characteristics including their domain knowledge, beliefs and perceptions about PF provision and epistemic emotions are likely to influence the development of such assessment skill. However, the roles of these individual characteristics are still widely ignored or not systematically addressed in research with preservice mathematics teachers. The aim of this dissertation is to investigate PF provision as one of the essential constituents of assessment for learning in mathematics education and to explore the role of students' individual characteristics in this learning activity.

1.1. Assessment for learning

Using assessment to assist learning is nowadays considered a necessity for every classroom at all levels of education. Such type of assessment is often referred to as *formative assessment* (see Black & Wiliam, 1998b; Black & Wiliam, 2009) or *assessment for learning* (Black, 1986). Some researchers do not make a clear distinction between the two terms (e.g., Black, Harrison, Lee, Marshall & William, 2004). Others, conversely, restrict the formative assessment term to carrying out assessment frequently during the teaching process without involving students in the assessment activities (e.g., Broadfoot, Daugherty, Gardner, Harlen, James, & Stobart, 2002). Assessment for learning (AfL) is defined in this dissertation as any practice by an agent (teacher, peer or student) that involves seeking and interpreting different sources of evidence about students' performance to evaluate and improve or modify the ongoing learning and teaching strategies resulting in progression towards pre-defined learning goals (Black & Wiliam, 2009; Sadler, 1989; Klenowski, 2009). In this dissertation, the term AfL is used because it is inclusive of (a) the active roles of the students and their peers, and (b) the role of feedback about what needs to be done to reach the learning goal(s) (see Wiliam & Thompson, 2007). Wiliam and Thompson (2007) suggested five key strategies that make assessment improve learning: (1) having clear learning goal(s) and success criteria shared with students and understood by them, (2) establishing classroom discussion and other sources of evidence to infer students' understanding, (3) providing feedback to students about how to reach their learning goal(s), (4) using peers as instructional resources, and (5) utilizing the role of students as actively responsible for their own learning. This dissertation focuses on feedback and the active role of students through investigating PF activities with preservice mathematics teachers.

1.1.2. Assessment for learning in mathematics education

Historically, mathematics was considered as knowledge that exists outside the person before the development of social psychology that shaped the perspective of regarding mathematical knowledge as a cultural practice (Edwards, Esmonde, & Wagner, 2011). The scientific community of mathematics education has established sociomathematical norms for mathematics classrooms that are different from classroom social norms (Yackel & Cobb, 1996). The more recent view of mathematics education acknowledges the active role of the students in constructing their mathematical knowledge. Since Black and Wiliam's (1998b) review, there were several recommendations to use AfL practices, particularly in mathematics. Wiliam published three articles (1999a, 1999b; 2000) discussing the importance of using AfL practices to raise the levels of achievement in mathematics. The articles focused on three main practices namely: rich questioning, feedback, and activating the student's role. Rich questioning provides insights into what students know and can do, and also provides teachers with evidence about what they have to do next to advance learning (Wiliam, 1999a). Feedback improves performance only if students utilize it, and that depends hugely on the quality of feedback (Wiliam, 1999b; Santos & Pinto, 2009). Activating the role of students in AfL activities can be achieved by sharing assessment criteria with students and involving them in self- and peer assessment (Wiliam, 2000). AfL activities are essential for preservice mathematics teacher-training programs, to prepare future teachers for implementing these activities in their classrooms. Several studies employed AfL practices including structured teacher feedback, and or self-assessment, peer assessment and co-assessment. These studies focused on improving students' mathematical competencies such as mathematical reasoning, modeling and problem solving skills (e.g., Balan, 2012; Gallimore & Stewart, 2014; Lauf & Dole, 2010; Jones & Alcock, 2014; Zerr & Zerr, 2011). What is shared by all of these studies is the emphasis on feedback or the active role of students in the learning process and the roles of these two AfL components in the development of mathematical competencies. However, studies with preservice mathematics teachers are scarce. In the next section, the role of students and the role of feedback in AfL are discussed in more detail.

1.2. The role of students in assessment for learning

Engaging students in assessment is an essential premise to AfL not only via considering their role in processing feedback information but also by involving them in peer- and self-assessment (Black & Wiliam, 2009; Broadfoot et al., 2002). The effective impact of peer assessment (PA) on learning has been reported by peer tutoring research (Graesser, D'Mello, & Cade, 2011; Topping & Ehly, 2001). Unlike teachers, peers are more often available and are closely involved in the learning experience since they are students themselves (Ladyshewsky, 2013). The availability of peers makes their feedback timely and more frequent as it can be easily provided on the spot while students are engaged in the learning task. While PA might not be as accurate as teacher's assessment, it is expected to help students in identifying the (small) next steps to improve performance (Topping, 1998). More specifically, "By engaging in dialogue about their knowledge and performance with peers, opportunities emerge for conversations about what they are learning and how this links in to their performance or knowledge base" (Ladyshewsky, 2013, p. 176). Furthermore, involving students in the assessment of their learning is expected to promote a sense of responsibility and ownership, and increased motivation (Topping, 1998).

1.2.1. Peer assessment and peer feedback

Peer Assessment (PA) is a learning activity where individuals or small group constellations exchange, react to, and (or) act upon information about their performance on a particular learning task with the purpose to accomplish implicit or explicit shared and individual learning goals. The configuration of PA can vary depending on the context it takes place in, the instructional design, and other interpersonal or intrapersonal characteristics likely to influence its products and outcomes. For example, PA can be conducted in many ways (e.g., unidirectional vs. bidirectional; in groups vs. in pairs; grades vs. written comments; once vs. iterative, etc.). PA can be regarded as a vehicle to self-assessment and self-regulation because students are likely to develop critique and assessment skills through the interactions with their peers in PA, and these skills may help them to assess their learning (Topping & Ehly, 2001). Topping (1998) proposed a typology of PA in higher education consisting of 17 variables on which PA varies: curriculum area, objectives, focus, product, relation to staff assessment, official weight, directionality, privacy, contact, year, ability, constellation assessor and assessee, place, time, requirements, and rewards (for more details about the typology see Topping, 1998). The typology was reclassified into clusters by some researchers (e.g., Van den Berg, Admiraal, & Pilot, 2006b; Van Gennip, Segers, & Tillema, 2009). Other researchers added new variables or redefined previous variables (e.g., Gielen, Dochy, & Onghena, 2011; Strijbos, Ochoa, Sluijsmans, Segers, & Tillema, 2009; Topping, 2010). For instance, in an attempt to align PA with collaborative learning Strijbos et al. (2009) proposed *interactivity* as a dimension of PA. Interactivity includes the directionality of PA (unidirectional or bidirectional), its frequency (once or more), and constellation (one to one, one to many or between groups). Additionally, Gielen et al. (2011) provided a more refined, updated 'checklist' or PA inventory for educators. This inventory was created based on a review of literature published after Topping (1998), which resulted in the addition of four new variables to Topping's (1998) typology and extending or redefining seven of the existing variables. Gielen et al.'s (2011) review produced a total of 20 dimensions of PA clustered into five main clusters: (1) decisions concerning the use of PA, (2) link between PA and other elements in the learning environment, (3) interaction between peers, (4) composition of assessment groups, and (5) management of the assessment procedure (for the extensive list see Gielen et al., 2011). Although Gielen et al. (2011) attempted to reduce the long list of the PA typology; they found that the diversity of PA research had increased since Topping's (1998) review. Despite the variety of PA activities, PF – the qualitative variant of PA – attracts many researchers because it fulfills the two key components of the AfL: (a) involving students in the assessment process and (b) providing feedback on performance. Another reason that contributes to PF as a favorable activity is that it can serve the separation of feedback from grades, a recommendation for better learning outcomes proposed by some researchers (e.g., Black & Wiliam, 1998a; Boud, 2000). Involving preservice teachers in PF activities can be beneficial for them, as they will need to provide feedback for their future students and peers. To provide a complete picture of PF, the role of feedback in AfL is addressed in the following section.

1.3. The role of feedback in assessment for learning

Feedback is widely regarded as an essential element to foster students' learning (Molloy & Boud, 2013). Therefore, it serves as a major component of AfL (Black & Wiliam, 2009; Sadler, 2010) as it is one of the mechanisms through which students become aware of their current performance, and how to reach their learning goals. Researchers often define feedback in terms of information about how successful students are in a particular learning task, but it can also be defined in terms of its effect (Sadler, 1989). The emphasis on the influence of feedback entails that the information should be regarded as feedback only if it is used by the student to close the gap between current performance and the learning goal(s) (Molloy & Boud, 2013; Sadler, 1989). Although Sadler (1989) already emphasized the importance of using the information, many feedback definitions in the literature do not explicitly stress this. For instance, Hattie and Timperley (2007) defined feedback as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). Similarly, Narciss (2008) considered feedback as "all post-response information that is provided to a student to inform the student on his or her actual state of learning or performance" (p. 127). Interestingly, even in the definition of *formative feedback* by Shute (2008) as “information communicated to the student that is intended to modify his or her thinking or behavior for the purpose of improving learning”. (p. 154), the application of information which makes feedback formative is not mentioned. It is important to note, therefore that information does not automatically trigger changes in students’ performance, as the students need to process the received information and decide whether to change their behavior/performance in response to that information. In PF, however, the composition of the information might be equally valuable as its implementation because the composition activity can be beneficial for the PF provider. In the next subsections, a brief overview of some of the prominent or more recent theoretical perspectives on feedback will be discussed, followed by a more detailed discussion of one feedback model and one feedback framework due to their relevance to PF practices. Then different factors influencing the effect of feedback on learning will be outlined, before presenting theoretical model/framework(s) of PA.

1.3.1. Theoretical perspectives on feedback

Several theoretical perspectives on feedback were proposed based on reviews or meta-analyses of research about the effectiveness of feedback. Some of the main theoretical perspectives include the feedback intervention (FI) theory (Kluger & DeNisi, 996), the Hattie and Timperley (2007) feedback model, the interactive tutoring feedback (ITF) model (Narciss, 2008, in press), and the interactional framework of feedback (Strijbos & Müller, 2014). All of these theoretical perspectives imply that the impact of feedback on performance depends on the type of processes being targeted by the feedback information, as well as factors related to the learning and the student. Whereas the Hattie and Timperley (2007) model provides guidelines to deliver feedback at different levels (task, process, self-regulation, and self), the FI theory (Kluger & DeNisi, 1996) presents theoretical assumptions about how feedback influences performance depending on the characteristics of the feedback (e.g., praise), the learning task (e.g., difficulty), and the student (e.g., level of self-efficacy). The ITF model

(Narciss, 2008, in press) and the interactional framework (Strijbos & Müller, 2014) stress the importance of internal feedback (i.e., internal processes used to judge own performance by the students themselves). This type of feedback makes it easier to understand how external feedback could be perceived and acted upon by the recipient. The most recent interactional framework by Strijbos and Müller (2014) acknowledges the role and the individual characteristics of the feedback provider, which are not explicitly addressed by the other theoretical perspectives, likely because most of them are based on teacher feedback research. Hattie and Timperley's model was found to be applicable to PF (Gan & Hattie, 2014; Harris, Brown, & Harnett, 2015). The Strijbos and Müller (2014) interactional framework can provide theoretical basis for the role of students in PF because it integrates the role of the PF provider and recipient in one framework. These two theoretical perspectives are further discussed below.

1.3.1.1. Hattie and Timperley's (2007) feedback model

Based on Sadler (1989), Hattie and Timperley (2007) proposed that feedback is about reducing the discrepancy between current performance and a learning goal (standard). They stressed that effective feedback should answer three questions: (1) "*Where am I going*?", (2) "*How am I going?*" and (3) "*Where to next*?" The first question is related to students' achievement of their goals (i.e., *"What are the learning goals?"*). The second question is about students' progress made towards the learning goals (i.e., *"What is the progress towards the learning goals?"*), and the third question is about further learning possibilities (i.e., "*What activities need to be undertaken to make better progress*?") (Hattie & Gan, 2011; Hattie & Timperley, 2007). As stated by Hattie and Gan (2011), for the first question (Where am I going?) students need to have clear goals and should have a clear understanding of their learning goals. Answering the second question (How am I going?) can involve a teacher, a peer, a task or the self, informing the student about current performance. The third question (Where to next?) can help in choosing better learning strategies (e.g., more self-regulated learning strategies, choosing different learning goals, different cognitive processes, deeper levels of processing, etc.).

According to Hattie and Timperley's model, the three questions can be answered at four different levels of feedback, namely task, process, self-regulation, and self. Task-level feedback is related to the product or outcome of the learning (i.e., whether the answer is correct or incorrect). This type of feedback is also referred to as corrective feedback. Process-level feedback is concerned with the learning processes used to perform the task (e.g., information search strategies, problem solving strategies). Whereas self-regulation level feedback is related to higher order processes such as monitoring the learning and regulating actions, self-level feedback involves evaluation of the self and the affect of the students (e.g., good girl!). According to Hattie and Timperley (2007), self-regulation level feedback is expected to stimulate engagement, leading students to involve in self-assessment and internal feedback.

Hattie and Timperley (2007) stressed that these different levels of feedback tend to have differential effects on learning. Self-level feedback is considered as the least effective level of feedback, whereas the other three levels are expected to have positive effects on learning. Self-regulation level feedback and process-level feedback are expected to have the most powerful effect because they stimulate deeper processing of the learning task compared to self-level and task-level feedback. Self-level feedback might have a detrimental effect on learning, because it diverts attention to the self and away from the learning task, and may lead to learned helplessness in some situations (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Self-level feedback is similar to what Kluger and DeNisi (1996) refer to as self-related meta-task processes (i.e., processes directing attention to the self, away from the learning task). The detrimental effect of self-feedback is attributed to it being personal, and having no information about the learning task. As stated by Hattie and Timperley (2007), excessive feedback at the task-level is, also, not very effective because it can distract the students and divert their attention to low-level processing ignoring the essential strategies required to reach the learning goals. The impact of the task-level feedback can also be "diluted" by combining it with some self-level feedback (e.g., praise) (Hattie & Timperley, 2007). The Hattie and Timperley (2007) feedback model will be revisited in Chapter 2 of this dissertation.

1.3.1.2. Strijbos and Müller's (2014) interactional framework

In an interactional framework, Strijbos and Müller (2014) integrate the role of individual characteristics and context into two distinct feedback processes: (1) the composition of feedback, and (2) the processing of the received feedback. Unlike other models that adopt a unidirectional perspective of feedback, in this framework both the roles of the feedback provider and the feedback recipient together with their individual characteristics (i.e., personal factors) in a specific context are equally acknowledged. This framework postulates that the processed (received) feedback message may be different from the provided feedback message because the individual characteristics of the recipient shape their interpretation of the feedback. Likewise, the individual characteristics of the feedback provider shape the way the feedback message is being composed. In the composition process, the feedback provider's individual characteristics, as well as their representations of the feedback recipient (i.e., how they perceive him/her), are activated, and the feedback message is composed in accordance with the interaction between these two feedback provider's factors. For example, a feedback provider who likes praise and perceives the recipient as being vulnerable to critique might provide a feedback message filled with praise. In the processing of feedback process, the recipient's reaction to the feedback message is a product of the interactions between: (a) the recipients' individual characteristics, (b) the recipient's representations of the feedback provider, (c) the characteristics of the feedback message, and (d) the recipient's *internal feedback*. To illustrate, a recipient who thinks that the feedback provider lacks the knowledge to provide effective feedback (i.e., representation of provider), receives vague feedback message (i.e., feedback message characteristics), and thinks that s/he already mastered the task (i.e., internal feedback), is likely to ignore the feedback message.

In general, the impact of feedback on performance is a product of the interaction between the characteristics of the feedback message, the characteristics of the learning context and the characteristics of the students. These factors are discussed next.

1.3.2. Factors influencing feedback effects

A major consensus in feedback literature is that providing feedback does not guarantee improved learning (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Sadler, 2010). Literature reviews and meta-analyses of feedback studies (e.g., Evans, 2013; Hattie & Timperley, 2007; Jonsson, 2012; Kluger & DeNisi, 1996; Shute, 2008; Van der Kleij, Feskens, & Eggen, 2015) all revealed contradicting findings of the positive impact of feedback on learning. Some feedback can even be detrimental to learning. Despite the importance of feedback to learning, students can ignore or not use the feedback information provided to them (Butler & Winne, 1995). Narciss and Huth (2004) proposed three main factors that influence the effect of feedback on performance which are related to (a) the feedback message characteristics, (b) the characteristics of the learning context, and (c) the student's individual characteristics. These factors are, also, likely to influence PF and are consequently discussed in detail next.

1.3.2.1. Feedback message characteristics

According to Narciss and Huth (2004), the impact of the feedback message depends on three main characteristics: (1) the functions of feedback (i.e., processes targeted by the feedback: cognitive, metacognitive, and motivational), (2) the contents of feedback (evaluative vs. informative), and (3) presentation of the feedback contents (e.g., timing, frequency, etc.). The functions of feedback are closely related to the instructional goals; whether they are information or motivational oriented (e.g., knowledge acquirement vs. reinforcing correct answers) (Narciss & Huth, 2004). The content of feedback consists of two components (evaluative or informative). These two components resemble the most common feedback types namely *outcome* (also referred to as verification) and *elaborated* feedback. The outcome feedback provides information about the learning product; whether it is correct or incorrect without providing any information about the processes involved in producing the outcome (Butler & Winne, 1995). Researchers frequently use variants of outcome feedback such as knowledge of results (KR), knowledge of performance (KP), and knowledge of the correct response (KCR) (Kluger & Denisi, 1996; Narciss, 2008). KR feedback provides information to the students about the correctness of their answers, whereas KP provides information about how well they performed in terms of the scores they achieved. In KCR feedback the correct answer is provided to the students (Narciss, 2008). Other types of less frequently used outcome feedback mentioned by Narciss (2008) are: answer until correct (AUC) feedback – which combines KR with the opportunity to revise work until the correct answer is reached – or having a limited number attempts to revise work in the case of multiple try feedback (MTF). Outcome feedback resembles task-level feedback in Hattie and Timperley's (2007) model. Since outcome feedback does not provide information to the students about how to proceed or what to do to reach the right answer (Butler & Winne, 1995), it therefore has less positive impact on learning compared to the elaborated feedback which provides such information to the students (Narciss, 2008; Shute, 2008). Elaborated feedback can have different targets; it can be about a response, task, discussion of error, providing guidance/prompts, and providing some examples (Shute, 2008). A third feedback type that is not related to the learning task and is mostly used for motivational purposes is *praise*, or what is referred to as self-level feedback by Hattie and Timperley (2007). Importantly, very complex feedback is assumed to impose high cognitive demand regardless whether it is an outcome or an elaborated feedback and may confuse the students and lead to lower performance (Hattie & Timperley, 2007; Kluger & Denisi, 1996; Shute, 2008).

The presentation of the feedback contents involves aspects related to the delivery of feedback such as frequency, timing, mode and format (Narciss & Huth, 2004). Feedback can be provided immediately after the student has completed a small step of a learning task, or it can be delayed until the student is done with the task. Findings about the impact of *immediate* versus *delayed* feedback appear to be mixed with field studies supporting immediate feedback and lab studies supporting delayed feedback (for a review see Shute, 2008). How frequent the feedback is delivered and whether it is written or verbal makes a difference. For instance, Van den Berg, Admiraal and Pilot (2006a) reported that feedback by students involved more comments about structure when it was written, but more comments on style when it was oral. Jolly and Boud (2013) suggested that oral feedback is more optimal in cases where frequent and timely feedback is required, for instance when students are working on a complex task or performing design or physical activities. In general, it is quite apparent that different feedback message characteristics (i.e., function, content, and presentation) are related the characteristics of the learning context and those of students, which will be outlined next.

1.3.2.2. Characteristics of the learning context

Many factors within the learning context influence the effect of feedback on performance, for example, the learning goals, the learning task, and sources of problems or errors (Narciss & Huth, 2004). Narciss and Huth (2004) stressed the importance of these three factors because the feedback message should be designed in accordance with them. It is well-established in the literature that teachers and students should have clear pre-defined learning goals, and the feedback should be crafted based on the learning goals (see Black & Wiliam, 2009; Sadler, 1989). The learning task is also central to the impact of feedback on performance, as it interacts with all of the other feedback and individual characteristics. The review by Shute

(2008) revealed that providing feedback on a complex learning task had a negative or no effect on performance. Furthermore, student's individual characteristics (e.g., domain knowledge, affect, and beliefs) can moderate the effect of feedback on performance when a complex learning task is used.

1.3.2.3. Student's individual characteristics

Student's individual characteristics that are suggested to influence the effectiveness of feedback include: domain knowledge, motivational-beliefs, academic self-efficacy, meta-motivational skills (Narciss & Huth, 2004), self-related beliefs (e.g., Miller & West, 2010), feedback related beliefs (e.g., Price, Handley, Millar, & O'Donovan, 2010), and emotions (Molloy, Borrell-Carrió, & Epstein, 2013; Ilgen & Davis, 2000). Since the processing of feedback can only be done by the recipient, no matter how clear and elaborated the feedback is, it will have no effect on student's learning if the student, for example, lacks the motivation to invest effort, or does not have sufficient domain knowledge to process the feedback. The impact of feedback on students' learning is expected to depend to a large extent on how that external information is filtered through students' pre-existing beliefs and motivation (Molloy, et al., 2013), which might trigger different cognitive and affective reactions. As a conclusion of their literature review on feedback provided in computer-supported instruction, Mason and Bruning (2001) suggested that the domain knowledge of students may influence how students benefit from different feedback and learning context characteristics. They proposed that low domain knowledge students may profit more from timely feedback when the learning task is easy or complex, but for high domain knowledge students delayed feedback may be more beneficial especially when the learning task is complex. Moreover, in a study by Miller and West (2010) elderly participants with high control beliefs were more engaged with a reading task after receiving positive feedback (i.e., verification of high-performance) compared to those with low control beliefs.

To summarize, students appear to benefit differentially from feedback depending on various individual cognitive, metacognitive, and affective factors. These individual characteristics interact with the characteristics of the feedback and the learning context to shape how students choose to process and act upon the feedback message. Potentially, this complexity may partly explain why there is no single magic recipe for effective feedback, and why students often ignore the feedback provided to them.

In teacher feedback models and studies, students are mainly addressed as feedback recipients. However, In PF students can be involved as providers and recipients, and their individual characteristics may have different effects on their learning depending on the role(s) of the students. The nature of PF necessitates the need for a process-based theoretical framework that illustrates PF as a PA activity and addresses the role of the PF provider, the PF recipient together with their individual characteristics, and the characteristics of the PF message. The next section discusses two theoretical perspectives on PA as well as an attempt to construct an integrative theoretical framework of PA that explicitly includes PF provision which is the focus of the current dissertation.

1.4. Theoretical perspectives on peer assessment

The diversity of PA practices (including PF) creates a need for a broad theoretical framework that addresses different cognitive, metacognitive, affective, and social processes of the assessor and the assessee. A multilevel model of PA skill is introduced first, followed by a multi-process framework of PA. Then both theoretical perspectives are integrated, and the role of PF provider and the recipient will be explicitly added to create a more inclusive model of PA.

1.4.1. The peer assessment skill model

This model was proposed by Sluijsmans and Prins (2006) based on a PA literature review and feedback from PA experts. According to this model, PA skill consists of several sub-skills namely, defining assessment criteria, judging the performance of a peer, and providing feedback for future learning. Each of these skills can be performed separately as part of the PA activity. The three skills consist of several activities that represent the second and third levels of the model (see Sluijsmans & Prins, 2006). Activities involved in defining assessment criteria are: (a) develop personal course objectives, (b) write a personal report on course objectives, (c) couple course objectives to study tasks, and (d) develop measurable criteria for each study task. Judging the performance of a peer involves analyzing the performance of that peer. Providing feedback for future learning consists of (a) formulate discrepancies in a peer assessment solution, (b) formulate suggestions for improvement, and (c) reflect on the suggestions for improvement to the peer. In a complete PA activity, students start by defining and creating assessment criteria, following which the performance of a peer is judged, and feedback for future learning is provided based on the observed peer performance (Sluijsmans & Prins, 2006). However, students can, also, only be involved in judging the performance of a peer, or in providing PF. Sluijsmans and Prins (2006) reported two empirical studies in teacher-training courses that showed the applicability of this skill model. This model supports the design and measurement of PA skills, as it decomposes PA into concrete activities which can be trained, practiced, and evaluated (Sluijsmans & prins, 2006). This model is, however, a skill model and therefore does not demonstrate processes and factors underlying each skill within the model. It also does not address the processes that students engage in as assessors and assessees in the PA activity, which is outlined in the process-oriented framework described in the next section.

1.4.2. The integrated multiple process conceptual framework of peer assisted learning

Topping and Ehly (2001) proposed a conceptual framework for *peer assisted learning* (PAL) activities (including PA) based on literature review. This framework relies on a socioconstructivist perspective on learning based on Piaget's (1959) and Vygotsky's (1930-1934/1978) theories (for a visual illustration of the framework see Topping & Ehly, 2001). Topping and Ehly postulated several theoretical assumptions about different processes that might be involved in PAL activities. Some of their assumptions, particularly those relevant to PA are summarized next.

According to Topping and Ehly's (2001) framework, five socio-cognitive and affective sub-processes feed into the main knowledge building process which involves the PAL processes. The first of the five sub-processes concerns the *organizational features* of the learning environment, such as learning goals shared by the students, time spent to give feedback on the task, and the degree of interactivity and immediacy. This sub-process is related to the characteristics of the learning context (i.e., learning goals, type of learning task, time for PA activity) proposed by Narciss and Huth (2004), and can be regarded as a part of the learning context in PA activities. The second sub-process refers to *cognitive conflict*, which is induced by the PAL activity. The third sub-process is *scaffolding and error management*, as the less able student needs support and scaffolding from the more able student within the zone of proximal development (Vygotsky, 1930-1934/1978). This sub-process can be supported through error detection, correction, questioning or through modeling depending on the type of PAL activity (Topping & Ehly, 2001). The fourth sub-process involves *communication* between the peers as a skill to be practiced and learned and as a means to execute the PAL activity. In any PAL, students need to communicate one way or another, and therefore the development of good communication skills is also central to PA. Topping and Ehly (2001) proposed that this sub-process emphasizes the role of language in developing thoughts (in alignment with the Vygotskian theory) because in a PAL activity, students can engage in listening, explaining, questioning, clarifying, simplifying, summarizing, hypothesizing, etc. They further suggested that these skills are important for carrying out PAL, but they can also enhance students' learning with practice. The fifth sub-process is *affect,* which involves students' motivation to engage in the task, having a sense of ownership, trust between peers, emotions triggered by assessing a peer or being assessed by a peer, etc. Some variables related to this sub-process investigated in PA research include psychological safety, value diversity, interdependence, and trust (see Van Gennip, Segers & Tillema, 2010).

According to Topping and Ehly (2001), all five sub-processes feed into the main process of extending each other's knowledge (declarative, procedural, and conditional). Extending knowledge (i.e., *knowledge building* process) is assumed to be achieved through improving current capabilities, modifying current knowledge, or adding new knowledge and building completely new understanding (Topping & Ehly, 2001). The knowledge building processes are expected to help the assessor and the assessee becoming more similar in their understandings (inter-subjective cognitive co-construction), which coincides with their perceptions of their learning and might not represent the actual learning (Topping & Ehly, 2001). In PA activities, however, the development of assessment is also regarded as a main outcome. This skill is sometimes more likely to be acquired through and/or improved via PA activities (for the assessor) compared to domain knowledge as suggested by Sluijsmans, Brand-Gruwel, van Merriënboer, and Martens, 2004, although there is evidence that lab reports writing skills could also improve as a result of multiple PA provision activities (Cho & Cho, 2011).

Topping and Ehly (2001) suggested that with practice and a higher frequency of PAL activities students can gain automaticity and fluency in carrying out PAL activities. With experience and instructional support, students are expected to become consciously aware of what happens during their learning interactions and monitor and regulate their learning strategies. Students' awareness then should optimally develop into metacognition which makes students more confident and self-assured that they can reach their learning goals (self-attribution and self-esteem) (Topping & Ehly, 2001). This process is akin to internal feedback (Narciss, 2008) and self-feedback (Butler & Winne, 1995), which is considered as the optimal goal of involving students in PA activities. These higher order cognitive and affective processes are assumed to feed back into the five sub-processes as students practice the PAL activity (Topping & Ehly, 2001). As stated by Topping and Ehly (2001), the degree to which each process or sub-process is likely to be utilized or achieved by students depends on the type of the PAL as well as the nature of support provided by the teacher to foster the outcomes of the PAL activities. It is important to note that this framework seems an idealized conceptual framework based on literature review of PAL research, and testing whether empirical evidence supports each assumption is beyond the scope of this dissertation.

1.4.2.1. Applying the framework to peer assessment

Topping and Ehly (2001) suggested that all of the five sub-processes can play a role in PA to varying degrees, and different processes in the model may be utilized depending on the design of PA. For instance, in PA of complex learning tasks, in which students have to provide elaborated PF and the assessment is reciprocal, interactive and iterative, students are expected to (a) experience cognitive conflict (especially in heterogeneous pairing), (b) provide frequent PF to each other, (c) engage in cognitive co-construction through which they reach a common understanding of the learning task, and (d) improve/modify their current understanding of subject matter (or procedural, conditional knowledge). Conversely, Topping and Ehly (2001) claim that PA activities that involve merely scoring and no discussion of the learning product are expected to utilize very few processes in the framework, and might not lead to cognitive, metacognitive, and affective gains.

1.4.3. Multi-level Multi-process Interactive Peer Assessment (M^{2}IPA) framework

The framework proposed by Topping and Ehly (2001) is inclusive of cognitive, metacognitive, and affective processes and outcomes, and takes the learning of the assessor and the assessee into account. Nevetheless, it operates at a meta-level, because it was proposed for various PAL activities (PA, peer tutoring, peer modelling, etc.). In contrast, Sluijsmans and Prins's (2006) model includes concrete PA activities (defining assessment criteria, judging peer performance, etc.), for which the framework by Topping and Ehly (2001) does not account. Depending on the activity performed in PA, whether it is providing PF or defining assessment criteria, several processes of Topping and Ehly's framework might be invoked. For instance, communication, organizing and engagement, and affect might operate during group discussions when defining assessment criteria. Through analyzing the study task and matching the course objectives with the study task, students might modify or build their knowledge. Likewise, they might engage in cognitive co-construction when devising assessment criteria that meet the learning objectives. Furthermore, when students are engaged in judging a peer's performance and providing PF, they might experience cognitive conflict and utilize several error management and communication processes (e.g., questioning, hypothesizing, summarizing), and might also reflect on their own learning.

Importantly, the Strijbos and Müller (2014) interactional feedback framework explicates the processes of composing feedback and processing of feedback – two essential processes not explicitly outlined in the framework proposed by Topping and Ehly (2001) and Sluijsmans and Prins's (2006) PA skill model – and constitute important components for an integrative PA model. These processes are assumed to take place while judging the performance of a peer and providing PF for future learning. Additionally, as suggested by Strijbos and Müller (2014), the composition and the processing of the (peer) feedback message is influenced by the individual characteristics of the provider/recipient (e.g., domain knowledge, self-efficacy, emotions, etc.), as well as the representation of the counterpart being either the recipient or provider of (peer) feedback (e.g., domain knowledge, vulnerability, trustworthiness, etc.). Similar to other processes within the Topping and Ehly's (2001) framework (e.g., knowledge building, self-monitoring), the composition and the processing of PF are assumed to have a reciprocal relationship with the sub-processes that feed into the PA activity (e.g., cognitive conflict, scaffolding and error detection, communication) (see Figure 1). The sub-process affect is one component of the individual characteristics of the provider/recipient in Strijbos and Müller's (2014) and fits better with factors likely to influence the PA activities. This sub-process could, therefore, be removed from the sub-processes proposed by Topping and Ehly (2001), as it is accounted for by the individual characteristics component.

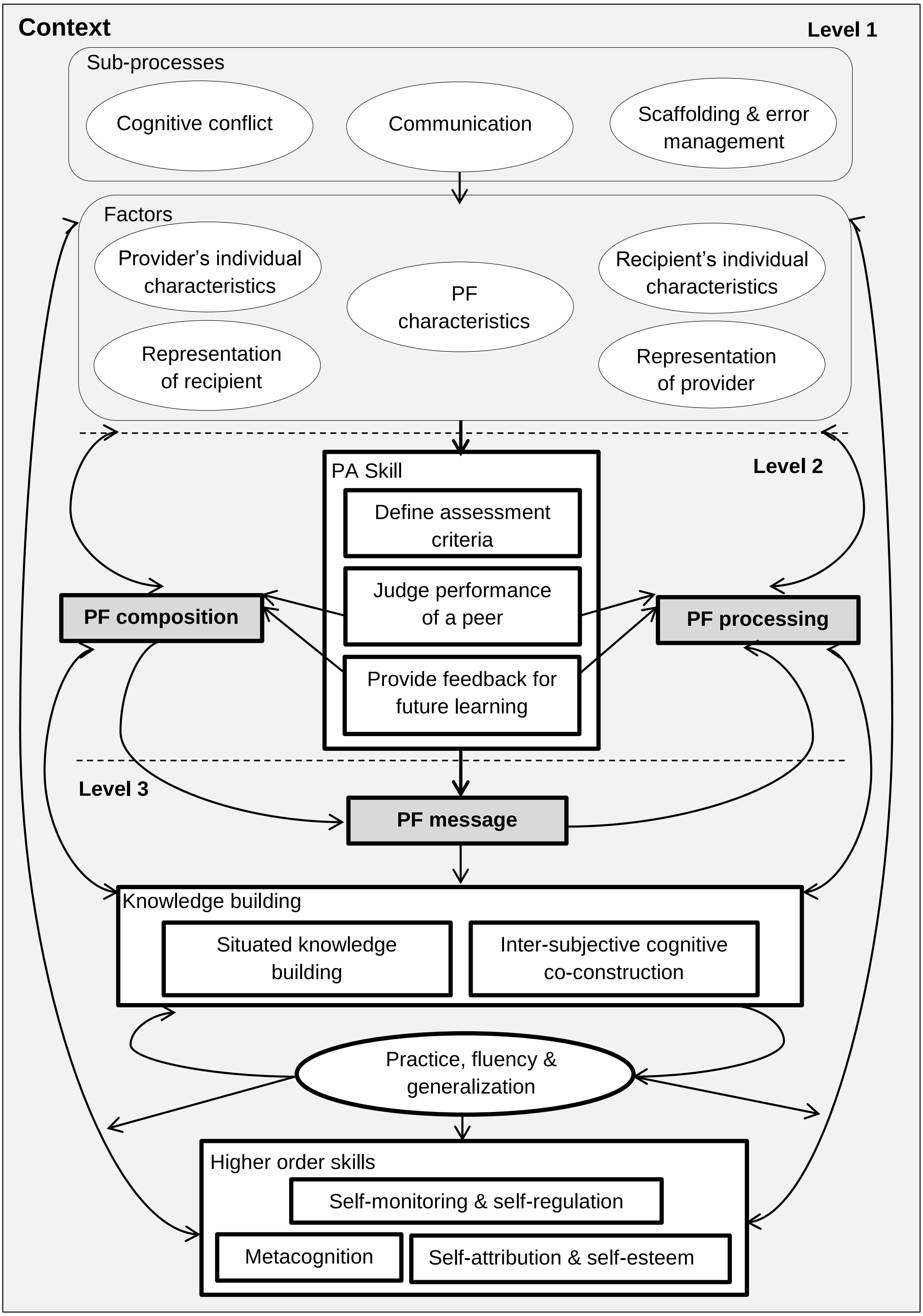

Figure 1. Multi-level Multi-process Interactive Peer Assessment (M²IPA) framework based on Topping and Ehly (2001) framework, Sluijsmans and Prins (2006) PA skill model and Strijbos and Müller (2014) interactional feedback framework.

Feedback literature, also, strongly emphasize the role of the feedback characteristics (e.g., Hattie & Timperley, 2007; Narciss & Huth, 2004), which can be expected to mediate the knowledge building process in PF activity. Empirical evidence suggested that the type of PF (e.g., specific or general) differentially effects PF recipients' performance (e.g., Lin, Liu, & Yuan, 2001; Cho & MacArthur, 2010), as well as PF providers' performance (e.g., Cho & Cho, 2011). It is, thus, assumed that the type of PF may influence PF composition, the processing of PF, and the degree to which some sub-processes (e.g., cognitive conflict or communication) are invoked. For example, a student who provides process-level PF or self-regulation level PF might be more likely to utilize communication sub-processes (e.g., elaboration, questioning) to resolve cognitive conflict compared to a student who provides self-level PF.

Integrating Strijbos and Müller's (2014) PF composition, processing processes and the PF message, as well as, the PF characteristics and the Sluijsmans and Prins (2006) PA skill model within the Topping and Ehly (2001) framework, provides a more detailed description of various processes that might take place in the PA activity. Additionally, Strijbos and Müller (2014) stressed the role of context in which different (peer) feedback processes take place (e.g., classroom, organization, informal setting, instructional support, etc.). Potentially, the processes and sub-processes within the Topping and Ehly framework and the PA skills model operate within a given context that shapes the learning experience of the assessor and the assessee. The learning context includes what Topping and Ehly describe as organizing and engagement sub-process (e.g., learning goals, time spent in the PA activity). Since this is the only sub-process that is related to the external learning environment and can be already represented by the context, it could be removed from the sub-processes.

The integration resulted in the "Multi-level Multi-process Interactive Peer Assessment" (M^2IPA) framework in which the type of PA activity students engage in, students' individual characteristics and PF message characteristics determine the degrees to which different processes and sub-processes are invoked within a specific learning context, and consequently the learning gains from PA (i.e., cognitive, metacognitive or affective) (Figure 1). The framework consists of three levels assumed to interact with each other within a given context. The first level consists of sub-processes and factors that can shape the PA activity, the second level consists of PA skill processes (e.g., defining assessment criteria, PF composition), and the third level represents the outcomes of the PA processes (i.e., PF message, knowledge building, and higher order skills). The exhaustiveness of the M^2IPA framework makes it difficult to test the entire framework empirically. Yet, it may serve as a guiding framework for empirical studies on PA. In the current dissertation the main focus is on PF provision; hence processes and sub-processes related to PF composition and the PF provider are expected to be activated during this PA activity. The next section provides a summary of empirical findings about effective PA practices.

1.5. Effective peer assessment practices

The effectiveness of PA in educational settings is often judged by comparing PA to teacher's assessment (e.g., Falchikov & Goldfinch, 2000). The outcomes of PA are often measured in terms of the validity of PA scores (compared to teachers' scores), improved performance (e.g., Cho & Cho, 2011; Gielen, Peeters, Dochy, Onghena & Struyven, 2010), improved PA skills, and/or attitudes and beliefs about PA (e.g., Sluijsmans et al., 2004; Strijbos, Narciss & Dünnebier, 2010; Van Zundert, Sluijsmans, & Van Merriënboer, 2010), and perceived learning (Van Gennip et al., 2010). Several literature reviews revealed that, so far, few studies measured students' learning gains as the vast majority of studies appear to be concerned with the PA validity (Van Zundert et al., 2010; Van Gennip et al., 2009).

A review of empirical PA studies by Van Zundert et al. (2010), revealed that: (a) training and practice are essential for PA accuracy, PA skills and attitudes towards PA, and (b) students' performance improves if they revise their work after PA. These findings support the assumption – adopted in the M^2IPA framework – regarding practice and fluency which are expected to result in better PA skill and knowledge building (Topping & Ehly, 2001). However, the findings tend to indicate that practice and fluency in PA is likely to be acquired via external support (e.g., training) and not necessarily through the PA involvement per se. Furthermore, in a recent meta-analysis of 69 PA studies, Li, Xiong, Zang, Kronhaber, Lyu, Chung and Suen (2015) reported a moderate average Pearson correlation between peer and teacher's assessment (.63). Factors reported to have a positive impact on the validity PA were: (a) PA is paper-based not computer-based, (b) students are involved in creating the assessment criteria, (c) subject is not medicine, (d) PA is done in higher education rather than in school education, (e) the assessment object is individual instead of group work, (f) matching of assessors and assessees is random, (g) PA is not anonymous, (h) PA is voluntary, and (i) PA involves qualitative comments in combination with scores (i.e., PF). Most notable in the context of the current dissertation is that validity is positively influenced by the use of qualitative comments, which is achieved through involving students in PF.

Since PF requires more than assigning a score, it is expected to involve more processes and/ or stronger influence of the sub-processes in the M^2IPA framework. For instance, in PF activities the composition and processing of PF are utilized which are likely to rely more on sub-processes including communication, cognitive conflict, and scaffolding and error management as compared to merely producing or receiving a score. As with any PA activity, PF can lead to gains for both the PF provider and the recipient depending on what the PF activity entails. The focus of this dissertation is mainly on the PF provider, therefore, the role and potential learning gains of the PF provider and their individual characteristics likely to shape PF provision activities are highlighted in the next sections, followed by a summary of research about learning from PF provision, and PF in mathematics education.

1.6. The role and gains for the peer feedback provider

PF providers can benefit from *learning through assessing,* because such activity requires them to think critically about the work of a peer and the learning task at hand. PF providers are assumed to exercise many cognitive processes and engage in cognitively demanding activities that help them consolidate their knowledge and deepen their understanding of the subject matter including questioning, making inferences, comparing and contrasting, analyzing and evaluating alternatives (King, 2002; Topping, 1998). While assessing the product of a peer, assessors can use many of the sub-processes and processes in M^{2}IPA framework. They may scaffold and monitor their peers, evaluate their performance, question peer's answers, elaborate on or summarize their comments, hypothesize about the understanding of a peer, and reflect on their own learning. Being actively involved in the assessment process may also stimulate a sense of responsibility, ownership or trust (Topping, 1998; Topping & Ehly, 2001). With practice, PF providers are assumed to develop assessment skills that help them become better at assessing and monitoring their own learning (i.e., external feedback is internalized as self-feedback; Butler & Winne, 1995). Despite the expected various cognitive, metacognitive and affective gains of the PF provider, so far few studies in mathematics teacher-training focused on training assessment skills through the use of PF provision or investigated individual, and PF characteristics influencing activities/processes involved in PF provision and this is one of the aims of this dissertation.

1.7. Research about learning from peer feedback provision

The learning gains of students who provide PF received less attention in PF research, although some studies showed that students writing improved by providing PF on lab reports in the domain of Physics (e.g., Cho & Cho, 2011; Cho & MacArthur, 2011). Additionally, training preservice teachers in PA skills was found to improve their PA skills (Sluijsmans et al., 2004). What remains unclear; however, are cognitive processing and the individual characteristics underlying such gains from PF provision. In a survey with university students, Nicol, Thomson and Breslin (2014) reported that during PF provision students engage in critical evaluation of a peer's solution and self-reflection, which makes them think critically about their own performance. However, as postulated by the M^{2}IPA framework, different sub-process and processes are expected to be utilized differently depending on the nature of the PF activity and the individual characteristics of students such as domain knowledge, emotions, beliefs and perceptions.

1.7.1. Role of domain knowledge, emotions, beliefs about peer feedback provision, and peer feedback providers' perceptions in peer feedback provision

The experience of providing PF is likely to be shaped by the providers' and the recipients' individual characteristics. Among these factors are domain knowledge, emotions, and students' beliefs (Narciss & Huth, 2004). Firstly, the review by Van Zundert et al. (2010) outlined domain knowledge as one of the influential factors on PA. Furthermore, the domain knowledge of the PF providers appears to be an important factor as two recent studies in the domain of academic writing revealed that the domain knowledge of the PF provider influenced the type of provided PF (Patchan & Schunn, 2015), and whether that PF was implemented by the recipient (Patchan, Hawk, Stevens, & Schunn, 2013). These findings highlight the importance of addressing the role of domain knowledge in PF research. Students' domain knowledge can, also, be assumed to influence how the sub-process cognitive conflict is utilized by the students, because the degree to which students experience cognitive conflict and manage to resolve it is likely to depend on their domain knowledge. The role of domain knowledge is expected to be more salient with complex problem solving tasks, because providing PF on these tasks is cognitively more challenging than providing PF on simple tasks (Van Zundert, Sluijsmans, Könings, & Van Merriënboer, 2012).

Secondly, although the evaluative nature of PF is likely to induce several emotions, the role of emotions is presently not systematically addressed in PA literature including PF studies. Epistemic emotions (e.g., curiosity, confusion, frustration, anxiety, excitement, boredom) that are triggered by cognitive conflict are not investigated yet in PF research to our knowledge. Anxiety is frequently reported as experienced by the students when involved in PF or PA (e.g., Cartney, 2010; Cheng, Hou, & Wu, 2014), however, its relationship with PF products or outcomes in not tested. Other epistemic emotions which might be experienced due to cognitive conflict during the PF composition especially when dealing with complex problem solving tasks (e.g., curiosity, confusion) are yet to be explored.

Thirdly, according to the theory of planned behavior, beliefs and perceptions serve as indicators of behavior (Ajzen, 2005). Therefore, students' beliefs about PF provision as a learning activity and their perceptions of their PF should be considered when involving students in such activities because these individual characteristics can serve as an indicator of how students approach the PF activity (e.g., their commitment, engagement, etc.). Investigating the beliefs of preservice teachers is particularly important because research showed that teachers' beliefs predict their classroom practices (e.g., Brown, Harris, & Harnett, 2012; Rubie-Davies, Flint, & McDonald, 2012). We need to understand preservice teachers' beliefs about PF in order to conduct teacher-training classes to help preservice teachers develop adaptive PF-related beliefs. Preservice teachers' domain knowledge, their epistemic emotions, their beliefs about PF provision and their perceptions of their PF are taken into consideration in this dissertation, because – as assumed by M^2IPA framework – these individual characteristics are likely to shape the processes and outcomes of PF provision. These individual characteristics will be addressed in detail in the empirical chapters of this dissertation (Chapters 2 and 3).

Combined, the reported findings highlight the importance of researching the PF provision process taking into account students' individual characteristics likely to influence this activity. Despite the growing number of PA and PF studies in mathematics education (e.g., Balan, 2012; Jones & Alcock, 2014; Lingefjärd & Holsquist, 2005; Lavy & Shiriki, 2014; Zerr & Zerr, 2011), quasiexperimental and experimental studies are still limited especially with preservice mathematics teachers.

1.7.2. Research about peer feedback in mathematics

Most of the findings about the impact of PA and PF in mathematics are based on qualitative studies. Several qualitative studies with school and university students claimed that using PA and PF can result in improved mathematics competencies (i.e., problem solving, mathematical modeling, mathematical reasoning) (e.g., Balan, 2012; Ernst, Hodge, & Schulz, 2015; Lingefjärd & Holsquist, 2005; Zerr & Zerr, 2011). The only quantitative study, we are aware of, compared PA scores of undergraduate mathematics students to teachers' scores, revealing a Pearson correlation of .77 (Jones & Alcock, 2014). Furthermore, the centrality of assessment of the performance of complex mathematical tasks that involve scientific reasoning, such as proofs has motivated some researchers to investigate the impact of PF on preservice mathematics teachers' assessment skills of these tasks (e.g., Lavy and Shiriki, 2014). This study, however, focused on qualitatively identifying what preservice mathematics teachers focus on when providing PF on peer solutions to geometry proofs. There is a need to further investigate PF provision by preservice mathematics teachers on such tasks because mathematics teachers are nowadays expected to be able to teach and assess mathematical argumentation in their classes (Selden & Selden, 2015b). Additionally, even students in schools are expected to be able to perform mathematical tasks that involve scientific reasoning and argumentation (National Council of Teachers of Mathematics, 2000). Scientific reasoning skills are described in the next section, before discussing the role of these skills in mathematics education and how PF can support the development of these skills.

1.8. Scientific reasoning and argumentation

Thinking scientifically is one of the 21st-century skills (National Research Council, 2012), which is considered as a main outcome of K-12 and higher education (Fischer et al., 2014). This skill is valued in many domains and is considered essential for everyday life (Dunbar & Fugelsang, 2004; Kuhn, 2010). Scientific reasoning can be defined in terms of mental processes employed with the purpose of knowledge seeking and knowledge enhancement (Kuhn, 2010). Scientific reasoning involves scientific activities (e.g., conducting experiments, interpreting data, etc.) and reasoning activities used frequently in sciences including induction, deduction, analogy, problem solving, and causal reasoning (Dunbar & Fugelsang, 2004). A discourse-based process that is closely related to scientific reasoning is argumentation, which involves taking positions and providing claims supported by reasons and evidence to justify the adopted claims (Chinn & Clark, 2013). Fischer et al. (2014) proposed that scientific reasoning and argumentation (SRA) in various domains (e.g., science, mathematics, medicine, psychology, etc.) can involve several epistemic activities which take place within different epistemic modes ranging from purely theoretical to pure practical goals (i.e., theory advancement oriented, science-based reasoning in practice, and artefact-centered). Fischer et al. (2014) identified eight epistemic activities in scientific reasoning: (1) problem identification, (2) questioning, (3) hypothesis generation, (4) construction and redesign of artefacts, (5) evidence generation, (6) evidence evaluation, (7) drawing conclusions, and (8) communicating and scrutinizing. The degree to which each epistemic activity is used can depend on the domain. For example, engineers are more likely to construct and redesign artifacts frequently; psychology students are often required to identify the problem when engaging in SRA, whereas in mathematics the problem is typically provided to the students. In the next section, SRA in the domain of mathematics is discussed in more detail.

1.8.1. Mathematics as a proving science

Reasoning and argumentation skills are central to mathematics education and should be integrated into mathematics instruction along with other core processes including problem solving, communication, connections, and representations (National Council of Teachers of Mathematics, 2000). Proofs are essential components of mathematics as a science, and indeed mathematics can be regarded as a proving science (Reiss, Hellmich, & Reiss, 2002). It is important to note, however, that mathematical reasoning is not limited to constructing proofs, but it is also required to make mathematical arguments about the validity of solutions of some mathematical problems (e.g., geometric construction tasks). In tertiary education, there is a strong focus on teaching students to comprehend and construct proofs (Zerr & Zerr, 2011). According to the standards of the National Council of Teachers of Mathematics (2000), students who reason mathematically should be able to: (a) recognize patterns, regularities and structure in real-world and mathematical cases, (b) investigate if those patterns occurred for a reason or accidently, (c) formulate and test mathematical conjectures, (d) and develop and evaluate mathematical arguments and proofs. A mathematical proof can be considered as a:

> coherent chain of argumentation in which one or more conclusions are deduced, in accord with certain well specified rules of deduction, from two sets of "givens:" (1) a set of hypotheses, and (2) a set of "accepted facts" consisting of either axioms or results that are known to have been proven true. (Schoenfeld, 1988, pp.157-158)

Boero (1999) proposed – an introspection based – expert mathematician's model of proving process, which shares similarities with the epistemic activities proposed by Fischer et al. (2014). The model involves two types of phases: an explorative inductive empirical-based phase (i.e., conjecture production), and hypothetical-deductive reasoning phase (i.e., mathematical proof construction). Several activities taking part in those two phases were identified by Boero (1999) including: (1) production of a conjecture which involves an exploration of problem, (2) formulation of statements based on textual conventions, (3) exploration of the formulated conjecture and testing its validity, identifying arguments and evaluating them in relation to theory and seeking relationships between different arguments (4) selectively developing theoretical arguments into a deductive chain (5) Formulating the chained arguments into a proof that meets the standards of the mathematics community and (6) reaching formal proof. As can be inferred from the Boero's model, several epistemic activities from the Fischer et al.'s (2014) framework are likely to take place during the formulation of formal mathematical proof (see Table 1). Although these activities are likely to be experienced by experts, students have to be instructed to engage in multiple epistemic activities when dealing with proofs in the academic context instead of treating proofs as merely given facts.

Geometry is considered as a useful area for students to develop reasoning and argumentation skills in mathematics (i.e., proving) (National Council of Teachers of Mathematics, 2000) because it involves figures and shapes with properties that allow the students to explore mathematical concepts visually (Schoenfeld, 1986). Geometry problems that involve argumentation can be in the form of proving tasks or construction tasks. As can be inferred from the Boero's (1999) model and Fischer et al.'s (2014) framework, proving involves several epistemic activities which might be challenging for students. There is ample evidence that students including preservice mathematics teachers experience difficulties with proofs or when asked to reason mathematically (Dreyfus, 1999; Kuzniak & Houdement, 2001; Healy & Hoyles, 1998; Reiss, Klieme, & Heinze, 2001; Tapan & Arslan, 2009). The difficulties encountered by students include proving, knowledge of proof methods, and proof validation (i.e., the act of judging the correctness of a purported proof) (Healy & Hoyles, 1998; Reiss & Heinze, 2001; Selden & Selden, 2003). In proof construction activities, students tend to use empirical arguments (i.e., measurement or evaluation of examples) instead of using deductive arguments (Schoenfeld, 1988), and they sometimes do not seem to understand the difference between an empirical argument and a deductive argument (Ghazan, 1993).

Table 1

Comparison between Boero's (1999) activities of mathematician's model of proving process and Fisher et al.'s (2014) epistemic activities of scientific reasoning

Boero's 1999 mathematician's model of proving process	**Fischer et al.'s (2014) epistemic activities of scientific reasoning**
Production of a conjecture which involves an exploration of problem	Problem identification
Formulation of statements based on textual conventions	Questioning and hypothesis generation
Exploration of the formulated conjecture and testing its validity, identifying arguments and evaluating them in relation to theory and seeking relationships between different arguments	Construction and redesign of artefacts, evidence generation and evidence evaluation
Selectively developing theoretical arguments into a deductive chain	Evidence generation and evaluation
Formulating the chained arguments into a proof that meets the standards of the mathematics community	Scrutinizing and communicating
Reaching formal proof	Drawing conclusions

Despite the overwhelming evidence that students struggle with performing mathematical proofs, some researchers suggested that students can learn to argue mathematically depending on the instruction styles employed (e.g., Jones, 2000). Geometry proof instruction often mainly involves presenting a correct proof by the instructor or passively illustrating how a geometric construction is performed (Schoenfeld, 1988). Involving the students in the processes of constructing or validating proofs is expected to be beneficial to students as it may stimulate active processing of the proof (Zerr & Zerr, 2011). One way to actively involve students including preservice mathematics teachers in proof validation is by engaging them in PF activities. The next sections will address the role of PF in SRA in general and in mathematics education more specifically.

1.8.2. The role of peer feedback in scientific reasoning and argumentation

The development of different sciences relies heavily on the evaluation of the scientific work produced by members of the scientific community (Osborne, 2010). Particularly in mathematics, a purported proof is only considered as a proof when the scientific community commonly accepts it as a proof (Manin, 2010). Scientists tend to search for weaknesses and limitations of any claim or argument proposed by other scientists (Osborne, 2010). In a learning society that embraces students' learning as a process and role akin to that of scientists, it is important to prepare the students for this role. By scrutinizing other people's ideas, an individual receives the opportunity to develop a better or a new understating of scientific phenomena. Some students are more inclined to reflect on their ideas, when they experience some cognitive conflict or conflicting arguments from other people in their learning environment (Osborne, 2010). Cognitive conflict in SRA between peers, which is also a sub-process in the M^2IPA framework, can be achieved by applying PF through the evaluation of evidence and claims (Fisher et al., 2014). However, it is important to note that students are not merely passive evaluators of their peers' work – especially when they provide PF (Nicol et al., 2014). The provision of PF is assumed to require careful scrutinizing of the provided claims and evidence by peers, identifying problems with their SRA, and constructing evidence-based arguments to support the PF message delivered to them. In mathematics, PF can be employed to engage preservice mathematics teachers in active validation (i.e., evidence evaluation) of geometry proofs or construction tasks, which may result in enhanced proof validation and proof comprehension.

1.8.3. Scientific reasoning in mathematics and peer feedback provision

The ability to evaluate the correctness of a mathematical proof (i.e., proof validation) is as important as comprehending a proof for preservice mathematics teachers. Students can be expected to more actively and critically read and validate mathematical arguments in a proof with some errors, as opposed to reading a fully correct proof (Zerr & Zerr, 2011). When they read proofs from textbooks, students know that the arguments in the proof are inherently correct, and they may, therefore, only read the proof at a superficial level (Hodds, Alcock, & Inglis, 2014; Zerr & Zerr, 2011). Proving involves several informal reasoning steps which are not included in the final proof version provided in the textbooks (i.e., the first five steps in Boero's model), that provide the final product of the proof instead of the processes that resulted in the proof (Reiss, Heinze, Renkl, & Groß, 2008). Therefore, presenting the correct proof should not be the only teaching strategy of proofs (Reiss & Heinze, 2004), especially for preservice mathematics teachers because the way they receive their instruction is likely to influence their teaching practices. PF can be regarded as a useful instructional tool to improve preservice mathematics teachers proof validation skills through critical evaluation of mathematical arguments. There is evidence that school students (in the fourth and fifth grade) became more aware of their mistakes in mathematics problem-solving after being involved in monitoring the reasoning by a fictional peer (Okita, 2014). Benefits of evaluating an erroneous peer solution, as suggested by Zerr and Zerr (2011), include learning to self-assess the correctness of their arguments (i.e, proof validation), and being exposed to a large diversity of proofs which can vary in their level of correctness, and might consequently lead to better proof comprehension.

Proof validation; however, seems to be a challenging task for high school and undergraduate students including preservice mathematics teachers. Studies with undergraduate mathematics students showed that they could not differentiate between correct and incorrect proofs (e.g., Inglis & Alcock, 2012; Selden & Selden, 2003; 2015b) and that they tend to focus on surface features (e.g., formulae) when validating a proof (Inglis & Alcock, 2012). Nevertheless, short self-explanation training was found to foster undergraduate students' comprehension of proofs as it directed them to focus on important aspects of the proof (Hodds et al., 2014). Based on these findings, we can assume that PF provision on geometry proofs and construction tasks might be challenging for preservice mathematics teachers because – as postulated by the M^{2}IPA framework – providing PF is expected to require more utilization of some sub-processes (e.g., communication: elaboration, summarizing, justifying, etc.) as compared to judging the correctness of a proof performed by a peer, or to comprehending a proof from a textbook.

Preservice teachers may require PF provision training that directs them to focus on important features of the peer solution while providing PF. The degree to which preservice mathematics teachers could benefit from PF activities may depend on the characteristics of the PF provider, the PF recipient, and the provided PF. Additionally, whether preservice mathematics teachers indeed engage in critical evaluation during PF provision when encountered with geometry proofs or construction tasks is still unclear. The processes and individual characteristics of the PF provider (e.g., domain knowledge, beliefs, and epistemic emotions) and the PF recipient (e.g., correctness/incorrectness peer solution) underlying benefits from PF provision are also still understudied. One way to understand and uncover the cognitive processing that preservice mathematics teachers engage in while providing PF on peer solutions to geometry proofs is through the use of eye-tracking. In the next section, the notion of eye-tracking research in educational context is introduced.

1.9. Eye-tracking in educational research

Monitoring the eye-movements of students is now widely used to infer cognitive processing in many domains (e.g., reading, mathematics, and science). Studies using eye-tracking rely on the so-called *eye-mind* assumption (Just & Carpenter, 1980), which states that what is being fixated on is cognitively processed. That is, the locus of attention tends to be closely related to the gaze location (Hyönä, 2010). In other words, the reader is likely to fixate on words that require further cognitive processing. The eye-mind assumption suggests that fixating on an object within the visual field occurs concurrently with cognitively processing that object, hence the gaze duration on an object or a word directly represents the time it takes to process that object or word (Just & Carpenter, 1976, 1980). In complex activities such as reading or problem solving, attention and eye gaze cannot be detached (Rayner, 2009).

Eye-movements differ depending on the nature of the encountered task, with more complex tasks requiring more or longer fixations (e.g., silent vs. loud reading or simple vs. complex problem solving) (Rayner, 1998). The most common measures of eye-movements are *fixations* and *saccades*. Fixations are defined as time periods (200-300 ms) in which the eye remains relatively still, and the information is processed from the visual field (Rayner, 1998; Rayner, 2009). Saccades are defined as “rapid movements of the eyes with velocities as high as 500° per second” (Rayner, 1998, p. 373). Cognitive processing takes place during fixations, whereas the role of saccades is to shift the position of the eye’s center (i.e., fovea) to a new location within the visual field for detailed analysis (Hyönä, 2010). Despite the immediacy of eye-tracking measures that makes it possible to measure different cognitive processes (Just & Carpenter, 1976), this method has its limitations. Eye-tracking has a large degree of inference because it does not provide any information about the success or failure of processing (Hyönä, 2010). A student might spend a long time reading a proof without being able to comprehend it fully. This issue creates a need for complementary measures such as performance or cued-retrospective reporting (Hyönä, 2010).

Eye-tracking methodology has been successfully used in different domains, mainly in reading (for reviews see Rayner, 1980; Rayner, 2009), mathematical problem solving (e.g., Hegarty, Mayer, & Green, 1992; Suppes, Cohen, Laddaga, Anliker, & Floyd, 1983) , proof validation (e.g., Inglis & Alcock, 2012), proof comprehension (e.g., Hodds et al., 2014), science problem solving (e.g., Tsai, Hou, Lai, Liu, & Yang, 2012), and multimedia learning (e.g., Canham & Hegarty, 2010; Meyer, Rasch, & Schnotz, 2010). Eye-tracking is a promising methodology to infer cognitive processing in PF activities, although the use of such methodology in PF research appears very limited. Only recently a study was published using eye-tracking to infer mindful-cognitive processing by students processing PF (Bolzer, Strijbos, & Fischer, 2015). In the context of PF provision on complex mathematical problems that require scientific reasoning and argumentation (e.g., geometry proofs and geometric construction tasks), eye-tracking can serve as a good measure of PF providers' focus of attention on different components of the peer solution they provide PF on. However, no study to our knowledge investigated this topic yet. To explore the potential of eye-tracking to infer cognitive processing during PF provision on peer solutions to geometry proofs, this technique was used in the empirical study reported in Chapter 3 of this dissertation.

1.10. Overview of dissertation: Studies about the role of peer feedback provision in validating geometric construction tasks and geometry proofs

Based on the issues discussed in the earlier sections, it seems that despite the possible usefulness of PF provision for preservice mathematics teachers' validation skills and comprehension of complex geometric tasks (i.e., constructions and proofs), PF is still not widely used in preservice mathematics teachers' instruction. Furthermore, the possibility to train PF skills is still not investigated with preservice mathematics teachers. Although individual characteristics were frequently emphasized in feedback models and frameworks as influential factors, some of these individual characteristics (e.g., domain knowledge, beliefs about PF provision, PF providers' perceptions of their PF, and epistemic emotions) are often not addressed explicitly in many PF studies. This dissertation attempts to address these issues in two empirical studies. In Study 1, we focused on training PF provision skills taking into account providers' domain knowledge, and changes in their beliefs about PF provision representing two important PF providers' individual characteristics in the MI^{2}PA framework. PF characteristics investigated in this study was PF content. In study 2, the PF composition processes (distinguished in the MI^{2}PA framework) was closely explored using eye-tracking methodology, and the relationships between PF content with other PF providers' individual characteristics were investigated including epistemic emotions, beliefs about PF provision, and perceptions of their PF. The impact of peer solution quality representing a major recipients' individual characteristic (i.e., domain knowledge) was investigated in this study. Figure 2 illustrates the elements of the MI^{2}PA framework that were investigated in the current project.

The first empirical study (Chapter 2) focused on training PF skills of geometric construction tasks, the role of domain knowledge, and beliefs about PF provision of preservice mathematics teachers. The second empirical study (Chapter 3) concerned the PF composition process by investigating cognitive processing during PF provision on different qualities of peer solutions to a geometry proof, and the role of preservice mathematics teachers' beliefs about PF provision, their perceptions of their PF, and their epistemic emotions. To address the main research aim in detail, more specific research questions were developed to be investigated in each study. Next, an overview of the chapters of this dissertation is provided.

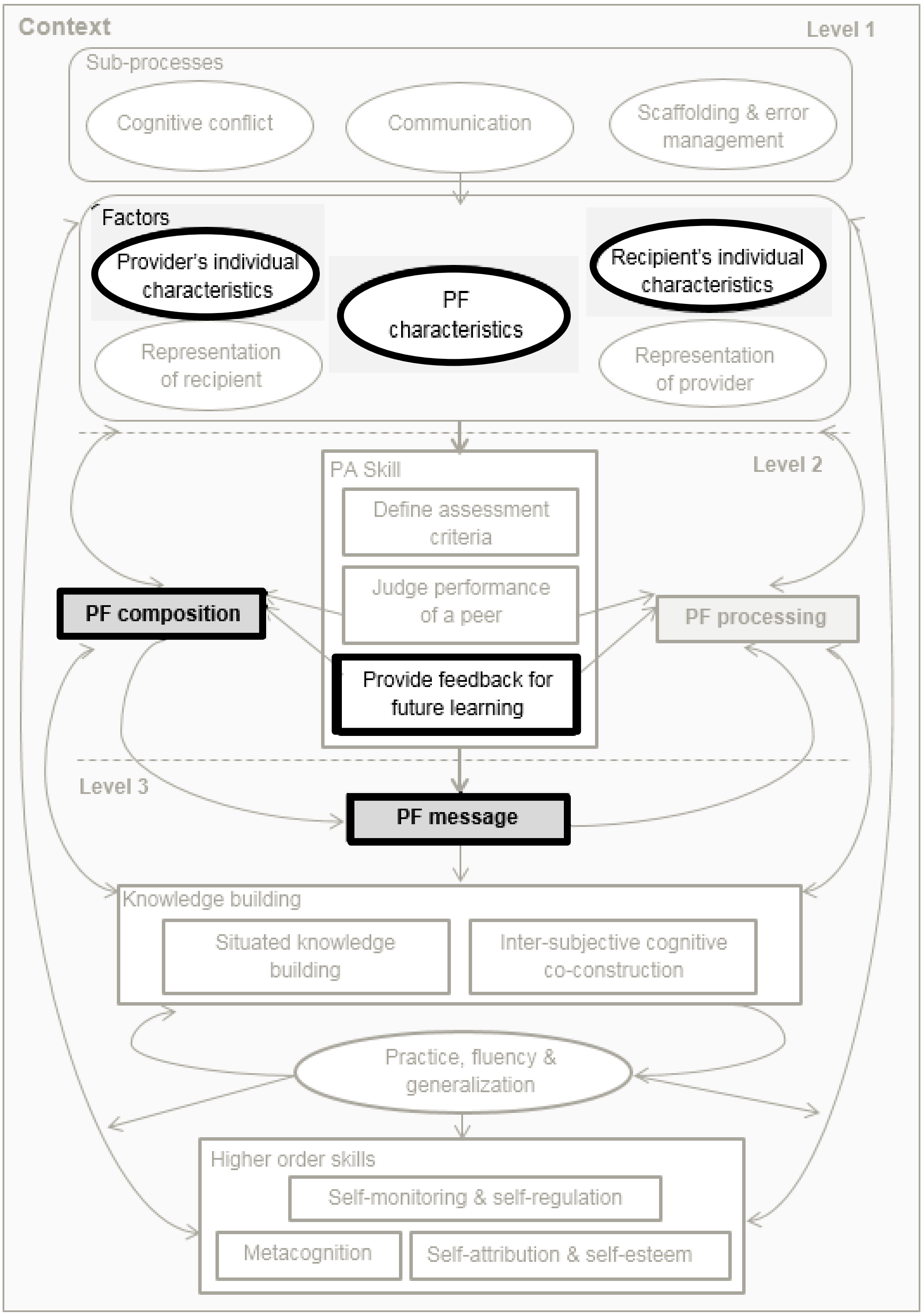

Figure 2. Illustrates the MI^{2}PA framework with the elements addressed in the current project highlighted.

Chapter 2 presents a study of training PF skills of 43 preservice mathematics teachers on geometric construction tasks. A quasiexperimental field study with a mixed design was conducted. All participants received PF training focused on Hattie and Timperley's (2007) feedback model, and they practiced providing PF on fictional peer solutions to geometric construction tasks over several sessions. Before the PF training, participants' PF skills, their basic geometric knowledge, and their beliefs about PF provision were measured. Performance on the basic geometric knowledge test was used to group the preservice mathematics teachers into three domain knowledge groups (low, medium and high). Participants' PF skills and their beliefs about PF provision were measured again after the PF practice sessions. The following research questions were investigated:

RQ 1. What is the impact of a structured PF provision training on preservice mathematics teachers' PF provision skills, and will students with different levels of domain knowledge benefit differentially from the training?

RQ 2. Will preservice mathematics teachers' beliefs about PF provision change after the training, and will domain knowledge play a role in that change?

Chapter 3 presents a study investigating preservice mathematics teachers' cognitive processing of peer solutions to a geometry proof during PF provision while taking into account their beliefs about PF provision, their perceptions of their PF, and their epistemic emotions. This study adopted an experimental between-subject design. The quality of fictional peer solution to a geometry proof was manipulated (near-correct vs. erroneous). Participants were required to provide oral PF to a fictional peer. Eye-tracking was used to monitor the participants' eye-movements while providing PF on the fictional peer solution to a geometry proof. After PF provision, the participants' proof comprehension, their basic geometric knowledge and their beliefs about PF provision, perceptions of their PF, and epistemic emotions were measured. Fifty-three preservice mathematics teachers participated, and the following research questions were examined:

RQ 1. How does reading cognitive processing while providing PF (measured by proportional total dwell time) differ depending on the quality of a peer solution (near-correct vs. erroneous) to a geometry proof?

RQ 2. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the PF providers' comprehension of the proof?

RQ 3. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the content of the provided PF?

RQ 4. What are the relationships between the PF providers' proportional total dwell time on the peer solution, PF providers' proof comprehension, and content of the provided PF?

RQ 5. What is the impact of the quality of the peer solution (near-correct vs. erroneous) to a geometry proof on the experience of the epistemic emotions curiosity and confusion?

RQ 6. What are the relationships between the PF providers' epistemic emotions, proportional total dwell time on the peer solution, proof comprehension, beliefs about PF provision, perceptions of their PF, and the content of their PF?

In the remainder of this dissertation, the two studies will be reported in more detail, followed by a general discussion in Chapter 4, in which general conclusions from both studies will be discussed. Finally, methodological limitations, research and practical implications, and directions for future research will be outlined.

2. Study 1: Training peer feedback skills on geometric construction tasks: Role of domain knowledge and peer feedback levels

2.1. Introduction

Providing feedback on the work of a peer is an essential analytical and communicative skill that is nowadays a requirement in many professional careers. One of the roles of higher education is to ensure that students are equipped with essential assessment skills which enable them to successfully make complex judgments about their own and their peer's performance in unstructured work environments (Boud & Falchikov, 2006). Assessment skills are particularly important for preservice teachers, as they will be assessing their future students and peers. Peer feedback (PF) fulfills the premises of the "assessment for learning" approach because (a) feedback is its core component and (b) students are actively engaged in the assessment process (Black & Wiliam, 2009). Thus far, PF research has hardly focused on the learning gains of the PF provider. Some studies showed that students' performance could improve after involving them in PF provision (e.g., Cho & Cho, 2010; Cho & MacArthur, 2011). However, engaging students in PF provision does not automatically result in learning. For instance, if students who lack domain knowledge focus only on surface features of peers' solutions (e.g., structure of a solution, spelling mistakes, etc.) it is unlikely that they will learn from providing PF. Domain knowledge is an important factor which is likely to influence PF provision and reception.

2.1.1. Role of domain knowledge in peer feedback research

The cognitive demand of a learning task can vary depending on the type and the difficulty of a learning task and the student's characteristics (e.g., domain knowledge) (Paas &

Van Merriënboer, 1994). This also applies to providing PF on a complex task, as it can be challenging for students with insufficient domain knowledge. A study by Van Zundert, Sluijsmans et al., 2012) showed that domain knowledge is a prerequisite for assessing the work of a peer, especially when the learning task is complex.

2.1.1.1. Peer feedback and complexity of the learning task

In most of the PF research, the objects of assessment were essays or reports. Few studies used complex tasks with scientific reasoning components (e.g., Gan & Hattie, 2014; Lavy & Shriki, 2014; Van Zundert, Sluijsmans et al., 2012), which are likely to stimulate higher order cognitive processing during PF provision. In mathematics education, geometric construction tasks are regarded as central to reasoning (Gulwani, Korthikanti, & Tiwari, 2011), but there is still limited evidence about the use of PF with such complex tasks. Further, since complex tasks are likely to induce higher order processing during PF provision, the role of domain knowledge becomes apparent for PF provision.

2.1.1.2. Domain knowledge and peer feedback provision

In any PF activity, the PF provider has to independently judge the correctness of a peer solution and to provide valid justifications for their judgment. Therefore, the type of provided PF is likely to be influenced by the PF provider's domain knowledge. Patchan and Schunn (2015) reported that, in academic writing, PF comments by students with low domain knowledge were dominated by praise, whereas comments provided by high domain knowledge students involved more criticism. However, there is still limited evidence about whether the type of PF provided on complex mathematical tasks like geometric constructions, is also influenced by domain knowledge.

2.1.2. Feedback levels

Due to the complex effects of feedback, several theoretical perspectives on feedback adopt a multidimensional (e.g., Narciss, 2008; Shute, 2008) or a progressive conceptualization of feedback (e.g., Hattie & Timperley, 2007). The Hattie and Timperley's (2007) model offers a conceptualization of student's engagement with the learning task and the learning and self-regulation processes associated with that learning task. According to this model, there are four types (i.e., levels) of feedback that have different effects on learning. The first level is feedback about the task (task-level), which refers the correctness/ incorrectness of the solution or the content knowledge used to solve the task. The second level is feedback about the learning processes and strategies that can be used to solve not only the task at hand but also other similar tasks (process-level). The third level is feedback about self-regulation, which directs the students to monitor and regulate their learning goals (self-regulation level). Self-regulation feedback does not provide information to the students about what they should do; instead, it stimulates students to act and reflect on their learning.

The fourth level of feedback is feedback about the self which includes no information about the learning task, but refers to personal characteristics of the student; most of the time in the form of general non-task related praise (self-level). This type of feedback is often reported in the literature as having a negative effect on performance (Kluger & DeNisi, 1996; Shute, 2008). Importantly, this feedback is not the same as what is referred to as self-feedback in the feedback literature (e.g., Butler & Winne, 1995). Self-feedback is internal feedback, which is part of self-regulation, whereas feedback about the self is personal and focuses on the self. For this reason, it is referred to as the self-level feedback.

Process and self-regulation feedback levels are claimed to be more beneficial for deeper processing and mastery learning because they stimulate deeper engagement with the learning task (Hattie & Timperley, 2007). Feedback at the self-level is regarded as the least useful feedback for learning because it diverts the student's attention to the self and away from the learning task (Kluger & DeNisi, 1996).

2.1.2.1. Feedback levels in peer feedback research

Due to its comprehensibility, several researchers used the Hattie and Timperley (2007) model to train or analyze PF of high school students in reading, mathematics, and chemistry (e.g., Gan, 2011; Gan & Hattie, 2014; Harris et al., 2015). These studies reported that PF was dominated by task-level and self-level PF (Gan & Hattie, 2014; Harris et al., 2015). Nonetheless, when prompted or trained to provide PF at the different levels, students could provide more PF at the process and the self-regulation levels (Gan, 2011). It seems that students can be trained to provide PF at higher levels (process and self-regulation), but it is unclear if this also applies to complex mathematics tasks such as geometric construction tasks.

2.1.3. Peer feedback and geometric construction tasks

Geometric constructions are useful scientific reasoning tasks because they are discovery tasks in which the results of learned mathematical problems (i.e., theorems or concepts) can be applied to physical objects in the real world (Schoenfeld, 1986). Similar to geometry proofs, constructions should be supported by deductions. However, research has shown that students at school and university levels including preservice teachers regard deductive proofs as irrelevant to geometric constructions (Kuzniak & Houdement, 2001; Miyakawa, 2004; Schoenfeld, 1989; Tapan & Arslan, 2009). When encountered with a geometric construction, preservice teachers tend to ignore the deductive mathematical knowledge they use to prove the object and employ an empirical approach using straightedge and compass (Kuzniak & Houdement, 2001; Tapan & Arslan, 2009). Preservice teachers should, therefore, be trained to use deductive reasoning when modeling geometric constructions to make this knowledge visible in geometry classrooms.

Engaging preservice mathematics teachers in PF provision in which they are instructed to provide PF on the process and self-regulation levels might be beneficial for their understanding of geometric constructions, as the provision of these types of PF requires deeper processing of the task. The progressive nature of the feedback levels in the Hattie and Timperley (2007) model can aid the preservice mathematics teachers to move beyond the surface features of the construction (i.e., how it looks). It may help preservice mathematics teachers to consider other processes through which the construction was created, as well as develop a deductive reasoning for the solution to be able to provide PF at the process and self-regulation levels. Nonetheless, providing PF at the higher levels (i.e., process and self-regulation) on geometric construction tasks can be expected to pose high cognitive demand, because this task is challenging for students to perform (Schoenfeld, 1989). Therefore, preservice mathematics teachers should be trained to be able to provide PF at the higher levels when evaluating these tasks.

2.1.4. Training peer feedback skills

For most students, including preservice teachers, PF is usually a new practice and students often experience uncertainty about their ability to assess the work of a peer and wish to have more support (Cheng & Warren, 1997). There is evidence that training preservice teachers to assess the work of peers results in better peer assessment skills (e.g., Sluijsmans et al., 2004). Moreover, training students with PF provision prompts was found to improve students' PF at the process and self-regulation levels (Gan, 2011). Thus, providing PF at the higher levels can be trained, but students might need some instructional support to help them deal with the complexity of some learning tasks such as geometric constructions.

2.1.4.1. Instructional support for peer feedback training

Van Zundert, Könings, Sluijsmans, and Van Merriënboer (2012) recommended training domain knowledge before training assessment skills when a complex learning task is used. However, in a typical classroom, face-to-face instruction time is usually limited, and sequential training of domain knowledge and PF skills might not be feasible. One way to account for this challenge is to scaffold students using domain knowledge and PF scaffolds. While the domain knowledge scaffold (e.g., worked example) provides a knowledge support for the PF providers, the PF scaffolds (e.g., prompts and evaluation rubrics) can direct them to focus their PF on certain aspects of the task. The combination of both scaffolds might ease the high demand imposed by the PF provision activity.

Worked example as a domain knowledge scaffold

Several studies in geometry and algebra education showed that teaching students using worked examples works better than the conventional teaching, even for students with low domain knowledge (e.g., Carroll, 1994; Paas & Van Merriënboer, 1994; Reiss et al., 2008). Using a worked example as a domain knowledge scaffold can be expected to reduce the demand introduced by the complexity of the task, and consequently, help students to focus on the PF provision process. Yet, the availability of the worked example does not necessarily guarantee that students will provide PF at higher levels (i.e., process and self-regulation), especially given that the majority of student comments tend to be about the correctness/incorrectness of the solution (Gan & Hattie, 2014; Harris et al., 2015).

Peer feedback provision prompts

Even in the presence of the worked example, providing PF at higher levels can be cognitively demanding, as students need to deeply process the peer solution to come up with PF at those levels. Structuring the PF activity through PF provision prompts can lead to more PF provision at the process and the self-regulation levels (Gan & Hattie, 2014; Gielen, Peeters, et al., 2010).

Evaluation rubric

Providing students with task-specific criteria (i.e., evaluation rubric) against which they judge the peer solution can help to focus on specific parts of the solution. The use of evaluation rubrics was found to increase the accuracy of peer assessment (Panadero, Romero, & Strijbos, 2013), and can be used in combination with the PF provision prompts to provide PF at different levels.

In summary, providing PF is beneficial to develop assessment skills, but preservice mathematics teachers have to be trained, and the PF provision activity should be structured to ensure that it is an effective learning activity. Although they received less attention thus far in PF research, students' individual characteristics – especially their domain knowledge – are essential for being able to provide PF. Other individual characteristics that can shape students' PF experiences are their beliefs about PF.

2.1.5. The role of students' beliefs in peer feedback

Student beliefs are central to PF research because they can predict how students approach the learning task (i.e., their behavior; Ajzen & Albarracín, 2007). In the area of PF, researchers appear to be concerned with changes in students' beliefs about PF given that they might be associated with students' insecurities regarding its usefulness and their ability to provide PF. Therefore, in many studies, students' beliefs about PF are measured before and after the study or intervention (e.g., Cheng & Warren, 1997; Sluijsmans et al., 2004), and these beliefs often change in a positive direction. In contrast, in a more recent EFL writing study, Wang (2014) reported a decrease in students' beliefs about the usefulness of PF which was attributed to several factors including, domain knowledge, limited writing skills, time constraints, and concerns about interpersonal relationships with peers. While domain knowledge is likely to play a role in how preservice mathematics teachers' beliefs about PF change, there seems to be no study to our knowledge that empirically tested this yet.

2.1.5.1. Beliefs about peer feedback provision and domain knowledge

Domain knowledge is not only a requirement for PF provision (Van Zundert, Sluijsmans et al., 2012); it can also shape preservice mathematics teachers' beliefs about PF provision. Preservice mathematics teachers who have sufficient domain knowledge are expected to have more confidence in their ability to provide PF, and are consequently expected to be more engaged and to believe that they can learn from providing PF. Those with low levels of domain knowledge might find providing PF overwhelming and, hence, adopt less positive beliefs about PF provision.

2.1.6. The present study

To summarize, although some studies used PF training to improve preservice teachers' evaluation skills, most of those studies were conducted in domains other than mathematics; typically, with written essays as the main PF object. Few studies investigated the gains of the PF provider when the learning task is complex (e.g., geometric constructions). Importantly, domain knowledge – which seems to be a prerequisite for PF provision – is still widely ignored. Despite the fact that several studies reported changes in students' beliefs about PF after being involved in PF activities, findings about the direction of change are mixed, and the impact of domain knowledge on such change is yet to be investigated.

2.1.6.1. Research questions and hypotheses

The aim of this study was to investigate the impact of preservice mathematics teachers' domain knowledge on their PF provision skills after practicing structured PF provision, and the impact of that experience on their PF provision beliefs. The research questions for this study were:

1. What is the impact of a structured PF provision training on preservice mathematics teachers' PF provision skills, and will students with different levels of domain knowledge benefit differentially from the training?
2. Will preservice mathematics teachers' beliefs about PF provision change after the training, and will domain knowledge play a role in that change?

Regarding the first research question, based on the findings by Gan (2011), it was hypothesized that after the training preservice mathematics teachers would provide more PF at the process and self-regulation levels (Hypothesis 1a). Based on the findings of Patchan and Schunn (2015), it was also hypothesized that particularly preservice mathematics teachers with high domain knowledge would provide more PF at the higher levels (process and self-regulation) after the training compared to those with low domain knowledge (Hypothesis 1b). Regarding the second research question, it was hypothesized, as found by Wang (2014), that preservice mathematics teachers' beliefs would change after the training (Hypothesis 2a), and that preservice mathematics teachers with low domain knowledge would have less positive beliefs about PF provision after the training (Hypothesis 2b).

2.2. Method

2.2.1. Participants

The participants were 58 middle school preservice mathematics teachers from a large university in southern Germany. Participation was a course requirement and the students received no additional compensation. The study ran throughout the semester. Data collection and PF training took place over several sessions. Only 43 out of the 58 students were present at the measurement sessions (9 males, 34 females, mean age 22.51, *SD* = 2.36) and were included in the analyses.

2.2.2. Design

In a quasiexperimental field study, a mixed design was implemented to investigate the quality of preservice mathematics teachers' written PF, and their beliefs about PF provision before and after PF provision training on geometric construction tasks, taking into account their basic geometric knowledge (see Figure 3).

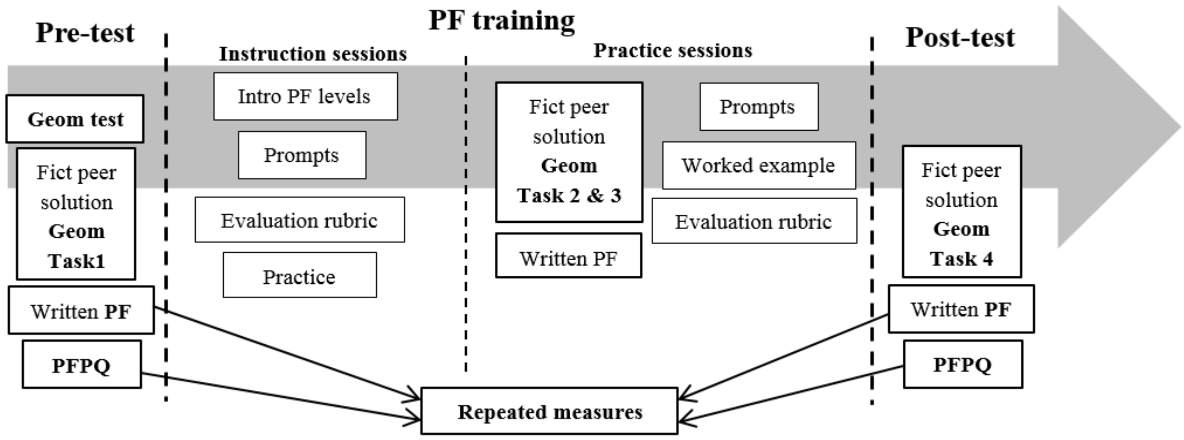

Figure 3. A detailed illustration of the study design showing different activities that took place during the peer feedback training in sequence. Geom test = basic geometric knowledge test; Fict peer solution = fictional peer solution to geometric construction task(s), Geom Task = geometric construction task; PFPQ= PF provision beliefs questionnaire; Intro PF levels = introduction to peer feedback levels.

2.2.3. Materials

2.2.3.1. Geometric construction tasks

The tasks presented preservice mathematics teachers with a set of geometric objects (e.g., a line, a circle, and an angle) and asked to construct – with ruler and compass – a specified object (e.g., a tangent to the circle that would intersect the line in a congruent angle). The task requirements were to (a) perform the construction, (b) describe their construction step by step, and (c) provide reasoning to show why their construction yields the specified object. When the construction task was introduced to the participants, they also received a basic geometric knowledge sheet that described basic definitions / rules and acceptable construction rules according to the mathematical norms of the teacher-training program. Fictional (peer) solutions were created for two construction tasks, one for the pretest and one for the posttest, based on pilot studies. Participants were instructed to provide written PF to a fictional peer about each part of the solution. All of the participants received the same peer solution. The fictional peer solutions contained (a) some correct steps, but partly followed an incorrect strategy, (b) correct descriptions of some but not all steps, (c) correct as well as incorrect reasoning steps, and (d) vague language in some parts of the solution. The graphical construction in the peer solution matched the description but was not performed with complete accuracy.

2.2.3.2. Training

All participants received PF provision training which consisted of two stages. In the first stage, two instructional sessions were held each lasting for 45 minutes. In the first instructional session, the notion of PF was discussed with the preservice mathematics teachers. The participants shared their thoughts about PF, its benefits, how it should look like, and their insecurities regarding PF. Then the feedback levels (task, process, self-regulation, and self) were introduced and discussed with the preservice mathematics teachers. At that point, the participants also received PF provision prompts accompanied with a task-specific evaluation rubric. In the remaining part of the first session and the second session, the preservice mathematics teachers were involved in several individual and group activities to understand the different levels of feedback better. They had to (a) identify each feedback level in written PF comments, (b) transform one feedback level to a higher level, and (c) work in groups to provide written PF on a solution and share and discuss their PF with the rest of the class. In the second stage of the PF training, which also involved two sessions, each participant received a fictional peer solution and practiced providing written PF on that solution with the help of the instructional scaffolds (i.e., prompts, evaluation rubric and worked example).

2.2.3.3. Instructional scaffolds

Participants received several instructional scaffolds that we introduced at different points during the PF training. In the first stage of the training, the preservice mathematics teachers received PF provision prompts and an evaluation rubric. These scaffolds were used in the instructional practice activities. In the second stage of the training, a worked example of each geometric construction task was provided in combination with the PF provision prompts and the evaluation rubric.

Peer feedback provision prompts

A visual organizer (developed by Gan, 2011) with progressive prompts reflecting different levels of feedback according to Hattie and Timperley's (2007) model was used (see Hattie & Gan, 2011 for the visual organizer). We extended Gan's visual organizer with additional prompts mostly knowledge integration/self-reflection prompts (adopted from Chen, Wei, Wua, & Uden, 2009; King, 2002; Nückles, Hübner, & Renkl, 2009) (see Table 2). Most of the added prompts were at the self-regulation level because it is assumed that the PF provider benefits more from providing knowledge integration/self-reflective questions as s/he needs to think deeply about the learning task (King, 2002).

Evaluation rubric

The evaluation rubric consisted of a set of criteria that could be used in combination with the PF provision prompts to judge the peer solution, and produce written PF. More specifically, the preservice mathematics teachers had to judge: (a) the construction of the geometric object, (b) the description of the construction and (c) the reasoning provided to prove the construction true.

Worked example

All participants received a standard worked example of the geometric construction task they had to provide written PF on in the second stage of the PF training (i.e., in the practice sessions). Since the purpose of this study was to improve preservice mathematics teachers' PF skills, and not their domain knowledge, we provided them with the worked example. This was done to ensure that those with low domain knowledge can still practice providing written PF at the higher levels (i.e., process and self-regulation) and not spend most of the sessions' time trying to understand the task. The geometric construction tasks used in this study always had more than one solution approach, and the solution proposed in the worked example was different from the fictional peer solution that the preservice mathematics teachers had to provide PF on.

Table 2

Prompts added to Hattie and Gan (2011) feedback levels visual organizer and their sources

Prompts	Peer feedback level	Source
Which parts were not sufficiently clarified?	Task	Nückles, Hübner & Renkl (2009)
How can you best explain…?	Self-regulation	
How would you use…. to…..?	Self-regulation	King (2002)
What would happen if you….?		
How does….tie to what we learned before?		
What conclusions can you draw about…?		
Explain why…..		
Explain how….		
How are …and … similar?		
How are… and ….different?		
What can you infer from….?	Self-regulation	Chen, Wu & Uden (2009)
What can you think of from….?		

2.2.4. Measures

2.2.4.1. Basic geometric knowledge test

To measure preservice mathematics teachers' domain knowledge, we used a basic geometric knowledge test used in previous research (Ufer, Heinze, & Reiss, 2008). The test consists of 49- true/false items. The test that measures different topics (e.g., properties of an equilateral triangle, properties of a parallelogram, transversals, etc.) was used at the pretest (Cronbach's α = .77) to group the preservice mathematics teachers' into the lowest, middle, and highest one-third of the sample.

2.2.4.2. Peer feedback provision questionnaire (PFPQ)

Participants' beliefs about: (a) learning from PF provision (LPF) (e.g., "I learn from providing PF"), (b) confidence regarding PF provision (CPF) (e.g., "I feel confident when providing positive feedback to my peers"), and (c) engaging in reasoning during PF provision (RPF) (e.g., "Providing PF helps me to be critical about my own arguments"), were measured before and after the PF training using the PFPQ. This questionnaire was developed to be used in the current study, and it consists of 40 items – five of which were adapted from Linderbaum and Levy's (2010) Feedback Orientation Scale. The items were scored on a 6-point Likert scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). Means were calculated for each subscale for further analyses.

The instrument (with initially 44 items) was piloted with an independent sample of students (N = 83). Parallel analysis following O'Connor's (2000) procedure revealed a three-factor solution, which corresponds to the theoretical structure of the questionnaire (Hinkin, 1998). The items were subjected to multiple PCAs, as the ratio of the sample size to the number of items was too small to include all of the 44 items in one analysis. Three rounds of PCA were conducted with two theoretically distinct scales as the components (LPF vs. CPF, CPF vs. RPF, and LPF vs. RPF) in each round. Items were only retained if they loaded meaningfully on the intended theoretical component with factor loadings > .40, and had a value of at least .50 for the diagonal elements of the anti-image correlation matrix (Field, 2009).

The procedure resulted in the exclusion of four items from the RPF scale. The three scales supported by the PCAs consisted of: (a) ten items for LPF (Cronbach's α = .87), (b) seventeen items measuring CPF (Cronbach's α = .91), and (b) thirteen items for RPF

(Cronbach's α = .88). PCA was not conducted for the current study as the sample size was not sufficient, but the scales' reliabilities were equally high for the present sample: LPF (Cronbach's α_{pre} = .83; Cronbach's α_{post} = .90), CPF (Cronbach's α_{pre} = .90; Cronbach's α_{post} = .96), and RPF (Cronbach's α_{pre} = 71; Cronbach's α_{post} = .87).

2.2.4.3. Peer feedback levels

Before coding participants' written PF, it was segmented following the procedure by Strijbos, Martens, Prins, and Jochems (2006), with the smallest meaningful segment as the unit of analysis. Two coders (first author and a student-assistant) independently segmented 10% of the data reaching an acceptable percentage agreement level (81.80% lower bound; 82.30 % upper bound), after which the first author segmented the remainder. The PF levels at the pretest and posttest were coded using a coding scheme based on Hattie and Timperley's (2007) model. The coding scheme was adapted from Gan and Hattie (2014). Although PF at the self-level (i.e., praise) is the least effective PF level and was not prompted for, it was included in the coding because PF often includes statements of this nature (Harris et al., 2015). The same two coders independently coded 10% of the segments (Krippendorff's α = .76). The first author then coded the remaining segments. Proportions of PF at each level (task, process, self-regulation, and self) were calculated for further analyses.

2.2.5. Procedure

In the first session, all participants completed the basic geometric knowledge test and the inventory on their beliefs about PF provision (i.e., PFPQ). Finally, they provided written PF to a fictional peer on a geometric construction task to measure their baseline PF skills. In sessions two to five, the participants received the PF training in which they learned to use the PF provision prompts and the evaluation rubrics and then practiced providing PF on two fictional peer solutions using the instructional scaffolds (i.e., PF provision prompts, evaluation rubric, and worked example). In the posttest session, each participant provided written PF on a fictional peer solution in the absence of all instructional scaffolds, and answered the PFPQ questionnaire again.

2.2.6. Analyses

Since PF provision using multiple feedback levels was trained, it was expected that the PF variables would not be normally distributed. Therefore, rank-based non-parametric ANOVA Type Statistic (ATS) tests for factorial designs (Brunner, Munzel, & Puri, 1999) was used to analyze the PF level variables using the R package nparLD (Noguchi, Gel, Brunner, & Konietschke, 2012). This non-parametric method does not require distributional assumptions and is claimed to be robust to outliers and small sample sizes (Noguchi et al., 2012). Instead of testing the hypothesis of equality of means, the ATS tests the hypothesis of equality of marginal distributions (Brunner, et al., 1999).

In the ATS tests for within-subject factors and interactions involving within-subject factors, the denominator degrees of freedom used for the approximation of the distribution is assumed to be infinity (Brunner et al., 1999). This is because the degrees of freedom used in conventional ANOVA produce conservative measures (Bathke, Schabenbergerm, Tobias, & Madden, 2009). A measure of the effect size for the ATS is the relative effect, which can be interpreted as the probability that a randomly chosen observation from the sample would result in a smaller value for a specific PF level (e.g., task or process) than a randomly chosen observation from a domain knowledge group y (e.g., low, medium or high) for a specific measurement occasion (i.e., pretest or posttest).

Changes in preservice mathematics teachers' beliefs about PF provision were analyzed using parametric ANOVAs. Although omega-squared (ω^2) is a less biased measure – compared to eta-squared – especially with small sample sizes (Okada, 2013), it cannot be used when the number of participants in each group is not equal (Field, 2009). Following recommendations by Lakens (2013), generalized eta-squared (η_G^2) is used as a measure of effect size for parametric ANOVAs. Values of .01, .06 and .14 indicate small, medium and large effects respectively (Cohen, 1988).

2.3. Results

2.3.1. Data inspection

The standardized skewness and kurtosis were within the acceptable range of ± 3 (Tabachnick & Fidell, 2013) for all PFPQ subscales, and there were no extreme outliers. However, the standardized skewness and kurtosis values were outside the acceptable range for the proportions of task-level PF (Skewness = -4.18, Kurtosis = 5.20), process-level PF (Skewness = 4.92, Kurtosis = 6.68), self-regulation level PF (Skewness = 7.82, Kurtosis = 13.63) and self-level PF (Skewness = 11.19, Kurtosis = 29.33) in the pretest; and for self-regulation level PF (Skewness = 3.49, Kurtosis = 3.28) and self-level PF (Skewness = 5.77, Kurtosis = 6.17) in the posttest. Seven univariate outliers were identified: one outlier for task-level PF, one for process-level PF, one for self-regulation level PF and two outliers for self-level PF in the pretest. One univariate outlier was identified for self-regulation level PF and one for self-level PF in the posttest. The outliers were checked to ensure that they were actual values. All outliers were retained and non-parametric tests were used to analyze the PF level variables.

2.3.2. Peer feedback levels after the training

To test for changes in the proportions of PF provided at each level (task, process, self-regulation, self) after the PF training, separate ATS tests were performed for each PF level with measurement occasion as the within-subject factor and students' domain knowledge (low, medium, and high) as the between-subject factor.

2.3.2.1. Peer feedback at task-level

There was a significant main effect for the change in the proportions of PF at the task-level, $F(1, \infty) = 4.07$, $p = .035$, and a significant main effect for domain knowledge, F(1.88, 34.87) = 3.95, p = .031. More importantly a significant interaction effect was found between domain knowledge and measurement occasion, F(1.86, ∞) = 4.99, p = .008 (Table 3, Figure 4a). Post-hoc multiple comparisons with Bonferroni correction revealed that low domain knowledge preservice mathematics teachers provided significantly more task-level PF after the training (M_{Rank} = 57.40, SD = 0.33) than medium domain knowledge preservice mathematics teachers (M_{Rank} = 23.93, SD = 0.31), F(1, ∞) = 11.78, p = .002. No significant differences were found after the training between low domain knowledge preservice mathematics teachers and high domain knowledge preservice mathematics teachers (M_{Rank} = 36.00, SD = 0.46), F(1, ∞) = 2.87, p = .273, or between medium and high domain knowledge preservice mathematics teachers, F(1, ∞) = 1.21 , p = .815.

2.3.2.2. Peer feedback at process-level

According to the ATS, the proportions of PF at the process-level did not change significantly after the training, F(1, ∞) = 0.45, p = .463 (Table 3, Figure 4b). There was no significant main effect of domain knowledge, F(1.95, 37.35) = 1.82, p = .177, and no significant interaction effect between measurement occasion and domain knowledge, F(1.99, ∞) = 1.65, p = .193. PF at the process-level did not significantly change after the training for low domain knowledge preservice mathematics teachers (pretest: M_{Rank} = 40.80, SD = 0.39, posttest: M_{Rank} = 34.53, SD = 0.31), medium domain knowledge preservice mathematics teachers (pretest: M_{Rank} = 44.61, SD = 0.46, posttest: M_{Rank} = 55.75, SD = 0.37), or high domain knowledge preservice mathematics teachers (pretest: M_{Rank} = 41.18, SD = 0.38, posttest: M_{Rank} = 44.69, SD = 0.40).

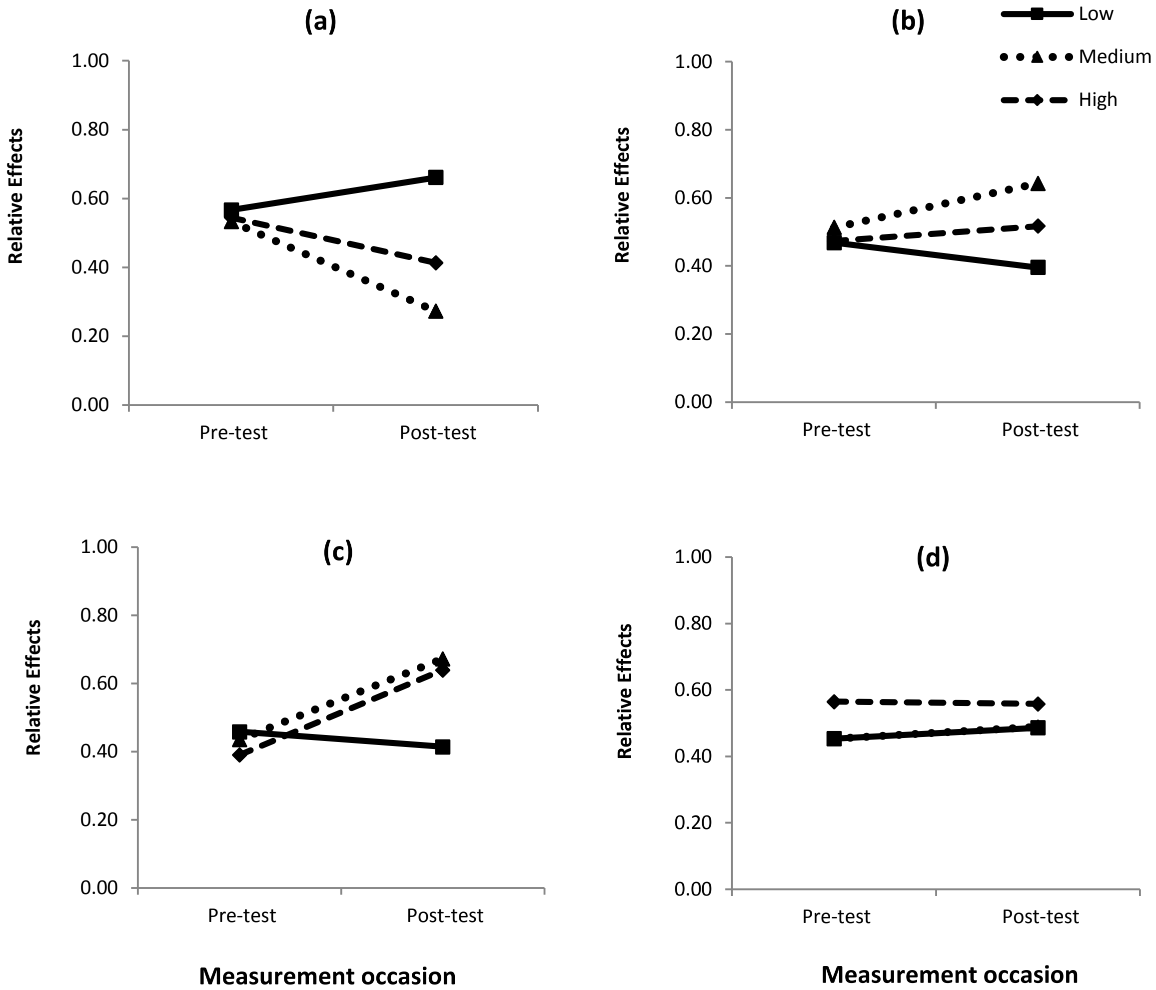

Figure 4. Relative effects of PF levels: (a) task, (b) process, (c) self-regulation, and (d) self, for each domain knowledge group (low, medium and high) before (pretest) and after (posttest) the PF training.

Table 3

Mean ranks and standard deviations of peer feedback levels (task, process, self-regulation, and self) before (pretest) and after (posttest) the training for each domain knowledge group

Peer feedback level	Group	*n*	**Pretest** M (*SD*)	**Pretest** 95% CI	**Posttest** M (*SD*)	**Posttest** 95% CI
Task	Low	15	49.20 (0.41)	[0.44, 0.68]	57.40 (0.33)	[0.56, 0.75]
	Medium	14	45.75 (0.40)	[0.41, 0.64]	23.93 (0.31)	[0.19, 0.38]
	High	14	47.32 (0.40)	[0.43, 0.66]	36.00 (0.46)	[0.29, 0.56]
Process	Low	15	40.80 (0.39)	[0.36, 0.58]	34.53 (0.31)	[0.31, 0.49]
	Medium	14	44.61 (0.46)	[0.38, 0.64]	55.75 (0.37)	[0.52, 0.74]
	High	14	41.18 (0.38)	[0.36, 0.59]	44.96 (0.40)	[0.40, 0.63]
Self-regulation	Low	15	39.90 (0.33)	[0.36, 0.57]	36.10 (0.25)	[0.34, 0.49]
	Medium	14	37.92 (0.31)	[0.35, 0.53]	58.32 (0.44)	[0.53, 0.78]
	High	14	34.04 (0.19)	[0.33, 0.45]	55.50 (0.39)	[0.51, 0.74]
Self	Low	15	39.50 (0.12)	[0.42, 0.49]	42.30 (0.21)	[0.42, 0.55]
	Medium	14	39.50 (0.12)	[0.42, 0.49]	42.50 (0.22)	[0.42, 0.56]
	High	14	49.07 (0.30)	[0.47, 0.65]	48.50 (0.28)	[0.47, 0.64]

2.3.2.3. Peer feedback at self-regulation level

The proportions of PF at the self-regulation level changed significantly after the training, $F(1, \infty) = 11.16$, $p = .001$ (Table 3, Figure 4c). There was no significant main effect of domain knowledge group, $F(1.90, 35.38) = 2.29$, $p = .118$. More importantly an interaction between domain knowledge and measurement occasion was found, $F(1.91, \infty) = 4.72$, $p = .010$. Post-hoc multiple comparison tests with Bonferroni correction showed that low domain knowledge preservice mathematics teachers provided less self-regulation PF after the training ($M_{\text{Rank}} = 36.10$, $SD = 0.25$) compared to high domain knowledge preservice mathematics teachers ($M_{\text{Rank}} = 55.50$, $SD = 0.40$), $F(1, \infty) = 9.02$, $p = .008$, and also less compared to medium domain knowledge preservice mathematics teachers ($M_{\text{Rank}} = 58.32$, $SD = 0.44$), $F(1, \infty) = 5.90$, $p = .045$. No significant difference was found between medium and high domain knowledge preservice mathematics teachers, $F(1, \infty) = 0.01$, $p = 1$.

2.3.2.4. Peer feedback at self-level

No significant changes were found in the proportions of PF at the self-level, $F(1,\infty) = 0.74$, $p = .388$ (Table 4, Figure 3d; note that lines for low and medium domain knowledge groups are superimposed in the figure). There was no significant main effect for domain knowledge, $F(1.30, 18.81) = 2.60$, $p = .117$. Similarly, no significant interaction effect was found between measurement occasion and domain knowledge, $F(1.82, \infty) = 0.32$, $p = .699$. PF at the self-level did not significantly change after the training for low domain knowledge preservice mathematics teachers (pretest: $M_{\text{Rank}} = 39.50$, $SD = 0.12$, posttest: $M_{\text{Rank}} = 42.30$, $SD = 0.21$), medium domain knowledge preservice mathematics teachers (pretest: $M_{\text{Rank}} = 39.50$, $SD = 0.12$, posttest: $M_{\text{Rank}} = 42.50$, $SD = 0.22$), or high domain knowledge preservice mathematics teachers (pretest: $M_{\text{Rank}} = 49.07$, $SD = 0.30$, posttest: $M_{\text{Rank}} = 48.50$, $SD = 0.28$).

2.3.3. Changes in Beliefs about peer feedback provision

Separate mixed-design ANOVAs were performed for each PF provision belief as the dependent variable, measurement occasion (pretest, posttest) as within-subject factor, and level of domain knowledge (low, medium, and high) as between-subject factor.

Table 4

Means and standard deviations of beliefs about peer feedback provision for each domain knowledge group before (pretest) and after (posttest) the training

		Low		**Medium**		**High**	
		n	M *(SD)*	*n*	M *(SD)*	*n*	M *(SD)*
LPF							
	Pretest	15	4.13 (0.75)	14	4.33 (0.48)	14	4.55 (0.57)
	Posttest		3.90 (0.82)		3.78 (0.99)		4.23 (0.64)
CPF							
	Pretest	15	3.74 (0.59)	14	3.84 (0.60)	14	4.03 (0.64)
	Posttest		3.28 (0.82)		3.54 (0.93)		3.68 (1.00)
RPF							
	Pretest	15	4.15(.38)	14	4.29 (0.61)	14	4.29 (0.49)
	Posttest		3.72 (0.48)		3.77 (0.97)		4.11 (0.52)

Note. LPF = learning from peer feedback provision; PCF = confidence regarding peer feedback provision; RPF = engaging in reasoning during peer feedback provision.

There was a significant decrease in learning from PF provision (LPF), $F(2, 40) = 10.93$, $p = .002$, $\eta_G^2 = .06$; in confidence regarding PF provision (CPF), $F(2, 40) = 9.81$, $p = .003$, η_G^2

= .06; and in engaging in reasoning during PF provision (RPF), $F(2, 40) = 26.19$, $p = .000$, $\eta_G^2 = .12$, (Table 4). There were no significant main effects of domain knowledge on preservice mathematics teachers' beliefs: LPF, $F(2, 40) = 1.23$, $p = .230$, $\eta_G^2 = .05$; CPF, $F(2,40) = 0.94$, $p = .400$ $\eta_G^2 = .03$; and RPF, $F(2,40) = 1.70$, $p = .20$, $\eta_G^2 = .06$. There were no significant interactions between domain knowledge and measurement occasion for LPF, $F(2,40) = 0.80$, $p = .457$, $\eta_G^2 = .01$; for CPF, $F(2,40) = 0.16$, $p = .851$, $\eta_G^2 = .002$; and for RPF, $F(2,40) = 0.33$, $p = .719$, $\eta_G^2 = .004$.

2.4. Discussion

This study investigated (a) whether preservice mathematics teachers, with different levels of domain knowledge, benefited differentially from structured PF training, and (b) how students' beliefs about PF provision changed in response to the training. Preservice mathematics teachers received training on providing PF at different levels (i.e., task, process, and self-regulation), and practiced providing written PF on two fictional peer solutions with the help of three instructional scaffolds: worked example, PF provision prompts, and an evaluation rubric.

2.4.1. Improvement of peer feedback levels

The results indicated an increase in PF provided at the highest level (i.e., self-regulation), but only for medium and high domain knowledge preservice mathematics teachers (Hypothesis 1a was partially supported; Hypothesis 1b was not supported). This finding suggests that engaging preservice mathematics teachers in PF provision activities structured with instructional scaffolds can help them process solutions to geometric construction tasks at deeper levels. However, domain knowledge appears to be a prerequisite as suggested by Van Zundert, Sluijsmans et al. (2012). This is also supported by the finding that low domain knowledge preservice mathematics teachers ended up providing more PF at the task-level after the training. Even when trained with a set of instructional scaffolds including a worked example, providing higher levels of PF seems to be challenging for low domain knowledge preservice mathematics teachers.

In contrast to the studies by Gan and Hattie (2014), and Gan (2011), PF at the process-level did not significantly increase after the training in our study. One explanation might be the type of task. In their studies, the object of PF was a chemistry lab report of an experiment which might elicit procedural comments to a larger extent. Conversely, in the case of geometric construction tasks many processes and strategies are rather implicit and therefore might be harder to provide PF on. Apart from the type of task, it might be due to differences in samples. Their study was conducted with high school students, whereas our study was conducted with preservice mathematics teachers who can be regarded as having more domain knowledge than high school students, and might not have realized the importance of providing process related PF to their peers. This difference requires empirical investigations by comparing the PF preservice mathematics teachers provide to that provided by high school students on geometric construction tasks.

Although no significant differences were found between medium and high domain knowledge preservice mathematics teachers in process and self-regulation PF levels after the training, medium domain knowledge participants tended to descripitvely provide slightly more PF at those higher levels. Further, only the medium domain knowledge group provided significantly less task-level PF than the low domain knowledge group after the training. These findings may suggest that medium domain knowledge students benefited most from the structured training in providing more PF at the higher levels. However, the relatively lower amount of PF at the process and self-regulation levels provided by the high domain knowledge preservice mathematics teachers does not necessarily indicate that they did not process the peer solution deeply. It might be the case that self-regulation level PF, which is closely related to internal feedback (i.e., self-feedback; Butler & Winne, 1995), has already been internalized by preservice mathematics teachers with high domain knowledge, and they might not consider it relevant to provide PF at that level. Alternatively, they might not be able to verbalize self-regulation PF as their mastery of performing the construction task might preclude them from verbalizing all individual steps of the procedure (Ericsson & Charness, 1994; Ericsson & Crutcher, 1991). Conversely, through the process of providing PF to a fictional peer, the medium domain knowledge preservice mathematics teachers – who still have room for improvement – might have realized the importance of the self-regulation PF for their improvement, and consequently used it more than their high domain knowledge counterparts. Further research with larger sample sizes is required to explore if such differences between high and medium domain knowledge with respect to the higher PF levels exist, and if so, find explanations for such differences.

2.4.2. Decrease in beliefs about peer feedback provision

Preservice mathematics teachers' beliefs about PF provision (LPF, CPF, and RPF) all decreased after the training with a medium effect size (Hypothesis 2a was supported). This finding is consistent with the study by Wang (2014) in which students' beliefs about the usefulness of PF decreased over repeated PF activities. Moreover, Wang (2014) suggested two related factors that contributed to this decrease in perceived usefulness lack of domain knowledge and domain skills. In the present study, this decrease was observed for all domain knowledge levels, so we can conclude that Hypothesis 2b was not supported. It might be the case that the geometric construction tasks were difficult for all participants and their beliefs became less positive regardless of their domain knowledge. Another potential explanation is provided by the self-assessment literature in which it is often reported that students who are low-performing on the target skill tend to over-estimate their performance (Panadero, Brown, & Strijbos, in press). Preservice mathematics teachers in this study might have over- or under-estimated their ability to provide PF due to the lack of or limited previous experience with PF.

When they were then introduced to the PF levels and repeatedly experienced producing PF at the higher levels on complex geometric construction tasks they might have realized that providing PF was more difficult than they expected. The decrease in beliefs observed in this study is inconsistent with previous studies in which students' beliefs became more positive after the peer assessment/ feedback activities (e.g., Cheng & Warren, 1997; Sluijsmans et al., 2004). Two factors appear to contribute to these contradictory findings: (a) the complexity of the PF object (i.e., essay vs. geometric constructions) (Van Zundert, Sluijsmans et al., 2012), and (b) the type of the PF product (i.e., grade, or feedback; low or high levels PF). Furthermore, while many PF studies – including our study – investigated changes in students' beliefs about PF provision, no study to our knowledge attempted to examine the impact of students' PF related beliefs on their performance, or on the type of PF they provide. Future research could examine (a) the impact of the complexities of the task and the PF product on students' beliefs about PF provision, and (b) the impact of different PF related beliefs on the PF type and learning outcomes.

2.4.3. Methodological limitations

Although the training improved PF provision at higher levels, separate effects of the instructional scaffolds could not be determined. Nevertheless, it is quite well established that training the domain knowledge of (relative) novices with worked examples works better than conventional teaching (for a review see Renkl, 2011) and that PF provision with the help of PF provision prompts works better than unprompted activity (e.g., Gan & Hattie, 2014). A study with larger sample size is required to vary different combinations of instructional scaffolds in PF training systematically, and then compare the content of PF between different conditions.

This study focused on PF provision skills with no direct measures of training effects on preservice mathematics teachers' task-specific performance (i.e., geometric construction task). While PF skills are essential for preservice mathematics teachers, it is important to investigate if training these skills can improve task-specific learning outcomes as well. Future studies could investigate whether PF provision training improves performance on construction tasks.

The progressive nature of the Hattie and Timperley (2007) multilevel feedback model makes it useful for training higher level (peer) feedback, yet it also creates grey areas between each level of feedback (e.g., task vs. process, or process vs. self-regulation). This can make it challenging to operationalize the different levels (e.g., coding scheme) and to apply them in practice (e.g., training purposes).

2.4.4. Practical implications

While it is often recommended in the literature that students with low domain knowledge can participate in well-designed PF activities, the findings of the current study show that domain knowledge is an important factor when it comes to the type of PF (i.e., progressively higher levels). Therefore, it is important to take into account preservice mathematics teachers' basic domain knowledge when designing PF training activities, as it might be challenging for preservice mathematics teachers with low domain knowledge to provide higher levels of PF even with the help of instructional scaffolds. Instructors should not assume that the instructional scaffolds automatically solve the problem of a lack of or low level of knowledge required to provide PF, especially for PF at the higher levels (i.e., process and self-regulation). A progressive training in which only domain knowledge is trained followed by PF skills training might be more beneficial for preservice mathematics teachers with low domain knowledge (Van Zundert, Könings et al., 2012). Finally, since preservice mathematics teachers' beliefs about PF provision change in response to repeated involvement in PF activities, instructors should carefully design PF activities to maintain positive perspectives about learning or being engaged in reasoning during PF provision.

3. The impact of peer solution quality on peer feedback provision on geometry proofs: links to peer feedback content, proof comprehension, and individual characteristics

3.1. Introduction

Involving students in assessment activities is considered one of the building blocks of preparing them to become lifelong learners (Boud & Falchikov, 2006). In mathematics education, one of the major roles of assessment is to help students become independent learners by being able to assess their performance (National Council of Teachers of Mathematics, 2000). *Proof validation* is considered as one of the key activities to proofs and proving instruction because it can help students to develop the skills to assess their own learning while performing proofs (Selden & Selden, 2015b). In such practice, students judge the correctness/incorrectness of a proof coming from different sources including proofs performed by a peer. Peer feedback (PF) is now frequently used as an activity that supports an active role of students. PF is the qualitative variant of peer assessment (PA) that is defined as a learning activity where individuals or small group constellations exchange, react to, and/or act upon information about their performance on a particular learning task with the purpose to accomplish implicit or explicit shared and individual learning goals. In mathematics education, PF only recently started to receive more attention (e.g., Balan, 2012; Lauf & Dole, 2010; Lavy & Shiriki, 2014). Being involved in PF activities is particularly valuable for preservice teachers because they need to develop assessment skills to assess the performance of their future students, their own performance and the performance of their peers.

Despite the growing evidence that students can benefit from providing PF (Cho & MacArthur, 2011; Cho & Cho, 2011), research investigating the underlying processes of PF provision is still limited. Instead, the focus of PF research has been widely on the characteristics of the PF message and how the PF message is being implemented by the recipient (Cho & MacArthur, 2010; Gielen, Peeters, et al., 2010; Guasch, Espasa, Alvarez, & Kirschner, 2013; Strijbos et al., 2010; Walker, 2015). Strijbos and Müller (2014) differentiate between the (peer) feedback message and the process of creating the (peer) feedback message (i.e., the composition process). According to their framework, the composition process is shaped by the (peer) feedback providers' individual characteristics (e.g., domain knowledge, beliefs, etc.), and the representations of the recipient (e.g., domain knowledge, vulnerability, etc.). For studies investigating the benefits of PF provision for preservice teachers, the composition process should be thoroughly examined to help developing more useful teacher-training programs. Additionally, less attention has been devoted to student's individual characteristics especially, domain knowledge, experienced emotions, other PF-related beliefs and perceptions and their roles in the PF composition process. PF provision on complex mathematics problems that involve scientific reasoning and argumentation like geometry proofs started to receive attention in the last years in teacher-training courses (e.g., Lavy & Shiriki, 2014). However, research exploring different processes underlying PF provision on peer solutions to geometry proofs while validating them is still limited. One way to investigate cognitive processing during PF composition on geometry proofs is through the use of eye-tracking methodology.

In the next section, research about learning from PF provision together with the benefits of providing PF on peer solutions to mathematical proofs will be addressed, followed by a discussion about PF content, and how some PF providers' individual characteristics such as domain knowledge, epistemic emotions, and beliefs/perceptions, could influence PF composition. A model of geometric reasoning to perform geometry proofs will then be introduced, with an attempt to extend that model to proof validation in PF provision. The notion of eye-tracking methodology based on Just and Carpenter's (1980) so-called *eye-mind assumption* will be finally introduced.

3.1.1. Peer feedback provision as a learning opportunity

Providing feedback on the work of a peer can be beneficial for the PF providers, not only because they become actively involved in assessment, but also because they can engage in several high-order thinking processes including, making inferences, questioning, generating hypotheses, weighing alternatives, and evaluating different information (King, 2002; Topping & Ehly, 2001). A qualitative study by Nicol et al. (2014) revealed that during PF provision, students reported engaging in a critical evaluation of their peer's and their own solution, experiencing self-reflection on their own learning, and generating explanations for their judgments. Another possible benefit of PF provision is being exposed to a wide range of solutions, which can enhance students' understanding of the learning task at hand (Zerr & Zerr, 2011). Studies with undergraduate physics students already showed that their writing skills improved due to providing PF on lab reports (Cho & Cho, 2011; Cho & MacArthur, 2011). Furthermore, PF training was found to improve preservice teachers' assessment skills (Sluijsmans et al., 2004). The development of assessment skills of some complex mathematical tasks such as proofs is essential for preservice teachers as they will need to teach and assess students' performance on these tasks in their future classrooms. Indeed, proof validation is considered as an important part of proof comprehension (Selden & Selden, 2003), thus involving preservice mathematics teachers in PF is expected to be beneficial for their proof comprehension and proof validations skills, but only if the PF activities align with the mathematical norms against which the proofs are normally being judged. Furthermore, in order to develop beneficial PF training for preservice mathematics teachers, factors, and processes influencing PF provision on peer solutions to mathematical proofs should be examined.

3.1.1.1. Peer feedback provision and validating mathematical proofs

One of the key elements to scientific reasoning and argumentation in mathematics education is mathematical proofs (Heinze & Reiss, 2001). Several proof-related activities are essential for proof teaching and learning including, performing a proof correctly according to the mathematical norms of the teaching context (i.e., *proof construction*), understanding a proof from a textbook or from lecture notes (i.e., *proof comprehension*), and judging the correctness of proof performed by students or by an external source (i.e., *proof validation*) (Selden & Selden, 2015a). Proof validation is an important skill for preservice mathematics teachers because school mathematics curriculum is likely to depend on proofs and justifications (Selden & Selden, 2015b). Previous studies revealed that proofs are challenging to validate for high-school and university students (Inglis & Alcock, 2012; Reiss, Heinze, & Klieme, 2000; Selden & Selden, 2003). There is growing evidence that when undergraduate students are asked to validate proofs of different levels of correctness they do not reliably differentiate between the correct and erroneous proofs (e.g., Alcock & Weber, 2005; Inglis & Alcock, 2012; Selden & Selden, 2003). Nevertheless, a recent study revealed that students' accuracy in proof validation depended on the type of error in the proof (Sommerhof, Ufer, & Kollar, 2016).

Geometry proofs are often used as initial context to teach scientific reasoning and argumentation in mathematics because they have a figure component that allows students to explore mathematical concepts visually and more easily by linking them to physical objects in the real world (Schoenfeld, 1986). Research with high school students (grades 7 and 13) showed that validating erroneous geometry proofs was more challenging to the students than validating correct proofs (e.g., Klieme, Reiss, & Heinze, 2003; Reiss et al., 2000). It is unclear, however, if preservice mathematics teachers also face the same challenge when validating geometry proofs.

When learning proofs, students are often exposed to formal proofs in textbooks or demonstrated by the instructor without showing the reasoning processes involved in creating that proof (Reiss et al., 2008). Since students are presented with correct proofs most of the time, they are likely to approach them passively and barely try to engage with a proof actively or to validate and comprehend its components (Zerr & Zerr, 2011). It is, therefore, not surprising that students have difficulties validating proofs because this process requires a deep understanding of how the proof is constructed (Selden & Selden, 2015a). Proof comprehension research showed that making the proof instruction more active by using self-explanation training resulted in improved proof comprehension (see Hodds et al., 2014). One of the major differences between proof comprehension and proof validation, however, is that in the latter activity the correctness of the proof is not quite evident (Selden & Selden, 2015a).

Involving students in PF provision is another approach to active engagement. When providing PF on a proof constructed by a peer, the PF provider engages mainly in proof validation. Yet, the PF provider also requires some comprehension of the correct parts of the proof to be able to identify possible errors and to provide PF accordingly. Some students consider comprehending the proof as a necessity to be able to validate the proof (Selden & Selden, 2015b). In one qualitative study by Zerr and Zerr (2011), PF provision activities on erroneous proofs were implemented to teach proof validation within a proof-based mathematics course at university level. The researchers reported that their students were descriptively more successful at identifying correct parts in the correct peers' solutions than being able to identify mistakes in the erroneous peers' solutions. What remains unclear, however, is whether the correct and erroneous proofs are processed differently by students when they are providing PF on them and whether students would comprehend the validated proofs differently. A study by Inglis and Alcock (2012) showed that when validating proofs, undergraduates adopt different reading behavior as compared to experts. Undergraduates were found to focus on the "surface features" of the proof (i.e., formulae), whereas experts focus on the logical structure of the proof. Nevertheless, no study yet to our knowledge has investigated whether preservice mathematics teachers cognitively process correct and erroneous peer solutions to geometry proofs differently when they provide PF on them and whether the quality of the peer solution influences preservice mathematics teachers' comprehension of the proof.

The conflicting findings from research about learning from erroneous worked examples in mathematics education (e.g., Große & Renkl, 2007; Heemsoth & Heinze, 2014; Isotani, Adams, Mayer, Durkin, Rittle-Johnson, & McLearen, 2011; Tsovaltzi, Melis, McLaren, Meyer, Dietrich, & Goguadze, 2010), together with findings from proof validation studies (e.g., Inglis & Alcock, 2012; Reiss et al., 2000) suggest that preservice mathematics teachers are likely to comprehend the proof they are providing PF on when the peer solution is not an erroneous proof. Yet, this assumption should be tested empirically. When investigating the processing of geometry proofs during PF provision, the PF content cannot be ignored because it can indicate what the PF providers attend to while providing PF on the peer solution to the proof.

3.1.2. Peer feedback content

According to the conceptual reviews and meta-analyses on feedback in education (e.g., Evans, 2013; Hattie & Timperley, 2007; Jonsson, 2012; Kluger & Denisi, 1996; Narciss, 2008; Shute, 2008), the effect of feedback on learning is very diverse, and can vary depending on the characteristics of the feedback message, the learner characteristics, and the learning context. Until recently, the characteristics of the feedback message received most of the attention in research on feedback and PF. One reason for that might be because the feedback message carries the information about current performance and directions for future improvement. However, the content of the PF message is perceived differently by PF researchers. For instance, Prins, Sluijsmans, and Krischner (2006) considered PF as a multi-dimensional construct measured by the use of assessment criteria, the nature of PF (e.g., remarks, question, suggestion), and the writing style (i.e., writing structure, formulation, and style). On the other hand, Van Steendam, Rijlaarsdam, Sercu, and Van den Bergh (2010) measured PF in terms of error detection, the accuracy of corrections of detected errors or suggestions made to change errors, and how correct, exhaustive and explicit the comments are.

Unlike experts, peers can lack domain knowledge, and therefore PF may include inaccuracies that are not necessarily represented in the PF writing style, use of assessment criteria, or nature of PF. This is particularly important for preservice teachers who will have to provide feedback for their students in the future, and their feedback should not only be of a particular type but should also be accurate. Accordingly, the PF content should be measured in terms of two dimensions, *type* and *accuracy*. Narciss and Huth (2004) identified three main characteristics of feedback type: (1) function (related to learning goals; cognitive, metacognitive and motivational), (2) content (evaluative or informative), and (3) presentation (e.g., frequency, timing, form, amount, etc.). Furthermore, PF can sometimes include non-evaluative general comments, comments about the writing style (i.e., surface features), or comments regarding the ability of the PF provider to evaluate the solution (i.e., self-efficacy; Strijbos, Van Goozen, & Prins, 2012). Strijbos et al. (2012) identify PF type in terms of its function (cognitive, metacognitive, affective, self-efficacy) and its style (evaluative / verification, informative / elaboration, or general).

In PF research, many studies seem to focus on (1) the type of PF provided by students (e.g., Patchan & Schunn, 2015; Walker, 2015), (2) training and/or scaffolding students to provide a specific type of PF believed to be beneficial for the recipient (e.g., Gan & Hattie, 2014; Gielen & De Wever, 2015;Gielen, Peeters, et al., 2010), and (3) investigating the impact of different types of PF on the recipients' revision of their work (e.g., Bolzer et al., 2015; Cho & MacArthur, 2010; Strijbos et al., 2010; Walker, 2015). Another strand of research compares the content of PF with teachers' feedback to investigate the accuracy of students' comments (e.g., Gielen, Tops, Dochy, Onghena, & Smeets, 2010; Hovardas, Tsivitanidou, & Zacharia, 2014; Patchan, Charney, & Schunn, 2009). Although these studies appear to focus on the PF providers by investigating their comments or training them to provide PF, the benefits gained by the PF providers from providing PF are still widely ignored. Furthermore, the relationships between the PF content, PF providers' cognitive processing of the learning task during the PF composition, and their individual cognitive and emotional characteristics are hardly addressed.

3.1.3. Role of individual characteristics: domain knowledge, beliefs and perceptions about peer feedback provision, and emotions

PF provision can be influenced by many students' individual characteristics among them domain knowledge, beliefs and perceptions about PF, and experienced emotions. Although PF providers' beliefs and perceptions about PF are more frequently addressed in PF research compared to domain knowledge and emotions, the relationships between these individual characteristics and PF content or performance are hardly explored. In the next sub-sections, the role of domain knowledge in PF provision will be first discussed followed by an outline of the role of beliefs and perceptions about PF, and finally, the role of emotions in PF provision is addressed.

3.1.3.1. The role of domain knowledge in peer feedback provision

Research revealed that domain knowledge is a pre-requisite for PF provision (Van Zundert, Sluijsmans et al., 2012), and it determines the type of PF that students provide (Cho & Cho, 2011; Patchan et al., 2009; Patchan & Schunn, 2015). In a study with high-school students, Van Zundert, Sluijsmans et al. (2012) found that for complex learning tasks domain knowledge becomes central for PF provision. Furthermore, Van Zundert, Könings et al. (2012) showed that sequential training of domain knowledge followed by PF provision skills is more beneficial to students than training them simultaneously.

The type of PF that students provide can differ depending on the domain knowledge of the PF provider and the PF recipient. Several studies showed that students with high domain knowledge produced PF that differed qualitatively from that provided by students with low domain knowledge (Cho & Cho, 2011; Patchan et al., 2009; Patchan & Schunn, 2015). Whereas students with high domain knowledge provide more criticism, their low domain knowledge counterparts provide more positive comments. Another factor that influences PF type is the quality of the peer solution. There is some evidence that low-quality written reports stimulated more elaborated PF comments, whereas high-quality reports stimulated more confirmatory PF comments (Cho & Cho, 2011; Patchan et al., 2009). Yet, a recent study by Patchna and Schunn (2015) showed that the only effect of the peer solution quality on PF type was the number of problems identified by high domain knowledge students. These mixed findings regarding the impact of peer solution quality on the type of PF have been all reported in the domain of writing research. The impact of the peer solution quality on PF content and its accuracy should be further investigated with other PF objects including mathematical proofs and with preservice mathematics teachers. Another PF providers' individual characteristic that can influence the content of PF and how preservice teachers engage in the PF activity is their beliefs and perceptions about PF.

3.1.3.2. The role of beliefs and perceptions in peer feedback provision

According to the theory of planned behavior, individuals' beliefs form the foundation of their intentions which in turns drive the individuals' behavior (Ajzen, 2005). Additionally, beliefs and perceptions are important interrelated psychological factors because they shape how the individual perceives the world and can also be shaped by the individual's repeated experiences (Smith, 2001). The Feedback Intervention (FI) theory postulates that not the characteristics of the feedback message per se that determine the effect of feedback on learning, rather the interaction between individual characteristics (e.g., personality traits, beliefs, perceptions, etc.) and the characteristics of the feedback message (Kluger & DeNisi, 1996). In feedback literature, several self-concept, (feedback) self-efficacy, motivational beliefs or perceptions were reported to be associated with specific individuals reactions to received feedback, feedback seeking behavior, and performance (e.g., Bernichon, Cook, & Brown, 2003; Fodor & Carver, 2000; King, Schrodt, & Weisel, 2009; Linderbaum & Levy, 2010; Williams & Johnson, 2000; VandeWalle & Cummings, 1997). Furthermore, some feedback related beliefs (e.g., the use of feedback) were found to be associated with (self-reported) self-regulation learning, academic self-efficacy, and GPA (Brown, Peterson, & Yao, 2016).

In PF research, the relationships between PF-related beliefs or perceptions, performance, PF content, and other learning related beliefs are hardly tested. A recent study by Brown et al. (2016) revealed that students' beliefs about PF (i.e., PF helps) were positively associated with (self-reported) self-regulation and negatively associated with GPA. Among the most commonly measured beliefs or perceptions of PF are 'usefulness' and 'fairness' of PF (e.g., Gukas, Miles, Heylings, & Leinster, 2008; Strijbos et al., 2010), confidence (e.g., Ertmer, Richardson, Lehman, Nrwby, Cheng, Mong, & Sadaf, 2010) and 'usefulness of PF for learning' (e.g., Ertmer et al., 2010; Sluijsmans et al., 2004). Qualitative studies revealed that students believed that they learn from providing PF, that they engage in reasoning during PF provision (Nicol et al., 2014) and that having sufficient domain knowledge influenced the usefulness of PF (Wang, 2014). According to the theory of planned behavior, perceived behavioral control such as self-efficacy can predict behavior (Ajzen, 2005). Therefore, it can be expected that PF providers' perceived control (i.e., PF self-efficacy, and sufficient domain knowledge) will be linked to the content of their provided PF. However, this assumption has not been tested empirically. Additionally, to our knowledge, no study measured PF providers' perceptions of their PF in terms of its quality, accuracy or usefulness, and tested the relationship between these perceptions and the content of provided PF. Investigating preservice teachers beliefs about PF provision is important because research showed that teachers' beliefs about teaching and learning predict their teaching practices (e.g., Brown et al., 2012; Rubie-Davies et al., 2011), and teacher training programs should be designed to foster adaptive beliefs about PF activities.

3.1.3.3. The role of emotions in peer feedback provision

Despite the important role emotions play in assessment and learning (Falchikov & Boud, 2007), research investigating different emotions experienced during PF activities is limited. There is evidence that the type of anticipated feedback can predict test-related emotions such as enjoyment, hope, pride, relief, anxiety, anger, hopelessness and shame (Pekrun, Cusack, Murayama, Elliot, & Thomas, 2014). Furthermore, when studying interactions in the classroom, anxiety has always been the focus of most researchers (Weiner, 2007); a condition that holds true for PF research as well. Students' lack of experience with PF can lead to experiencing some anxiety, which might be due to perceiving PF as unfair or due to lack of confidence to evaluate the work of a peer (Cartney, 2010; Hou & Cheng, 2012). Exposure to PF was reported to result in reduced test anxiety (Sluijsmans, Brand-Gruwel, & Van Merriënboer, 2002). Several studies investigated emotional perceptions in response to different PF practices (Praver, Rouault, & Eidswick, 2011; Raes, Vanderhoven, & Schellens, 2013; Strijbos et al., 2010; Vanderhoven, Raes, Montrieux, Rotsaert, & Schellens, 2015). Students' feeling of more comfort during PF provision was found to be related to anonymity (Raes et al., 2013; Vanderhoven et al., 2015). Students reported being more embarrassed and nervous when they had to evaluate the work of peers using scores and written comments compared to giving scores without comments (Praver et al., 2011). In addition, a previous study that measured students affect (offended, satisfied, angry, confident, frustrated, and successful) in response to PF reception, showed that students experienced more negative affect when they received elaborated specific PF from a high domain knowledge (fictional) peer, or when they receive general concise PF from a low domain knowledge (fictional) peer (Strijbos et al., 2010).

All of these studies examined social- or confidence-related emotions, but epistemic emotions likely to be experienced when PF providers engage in cognitive processes to comprehend or to validate peer solutions have not yet been studied.

In the PF composition process, students might experience cognitive conflict when they encounter peer solutions that differ from their own solutions or arguments they do not support. According to Pekrun and Stephens (2012), students experience some emotions (a) as a result of being cognitively involved in a task or (b) when they encounter cognitive conflict. Cognitive conflict triggers epistemic emotions such as curiosity when encountering a different approach to solve a problem (Kang, Hsu, Krajbich, Loewenstein, McClure, Wang, & Camerer, 2009), confusion when the conflict is not resolved (D'Mello & Graesser, 2012), enjoyment when the problem is solved, and frustration or anxiety when the cognitive conflict cannot be resolved (Pekrun & Stephens, 2012). When providing PF on a peer solution to geometry proof, curiosity, confusion, and anxiety are expected to be influential. Whereas, curiosity was found to positively predict shallow and deep learning strategies, confusion negatively predicted these strategies during mathematics problem solving (Muis, Psaradellis, Lajoie, Di Leo, & Chevrier, 2015). When students have to provide PF on a proof constructed by a peer, they might experience curiosity to figure out how the proof was approached, confusion when parts of the peer solution are contradicting with their own understanding, and anxiety when they cannot comprehend the proof and cannot provide accurate PF. The quality of the peer solution is expected to influence how different epistemic emotions are experienced. An erroneous peer solution containing contradicting information with PF providers' understanding might trigger more curiosity and confusion than a near-correct or a correct peer solution that is unlikely to trigger cognitive conflict. Furthermore, experiencing anxiety due to failure to resolve cognitive conflict is likely to be associated with failure to provide elaborated or accurate PF. To date, epistemic emotions experienced by preservice mathematics teachers during PF provision on peer solutions to a geometry proof, and the differential effects of the quality of peer solutions on experienced curiosity and confusion, as well as, the relationship between curiosity, confusion, anxiety and PF content have not been explored. Furthermore, epistemic emotions are triggered by the cognitive properties of the learning task (Pekrun, Vogl, Muis, & Sinatra, 2016), and these properties can also determine the type of reasoning used by students.

3.1.4. Geometric proofs reasoning through mental models

Performing geometric proofs is regarded as a complex task that usually involves multiple qualitatively different phases (Koedinger & Anderson, 1990). A mental-model based framework which explains some reasoning processes underlying geometric proofs construction was proposed by Ufer, Heinze, and Reiss (2009). This framework extended the Mental Model Theory (MMT) of deductive reasoning – according to which deductive reasoning is performed based on constructing and testing mental models (Johnson-Laird, Byrne, & Schaeken, 1992) – by adding a fourth phase that is specific to mathematical reasoning. Mental models are internal representations of premises or perceptual information in the external world that can be in the form of pictures, strings or symbols (Johnson-Laird, 1983; Johnson-Laird et al., 1992). The MMT postulates that deductive reasoning involves three phases. First, a mental model is created based on perceived verbal or perceptual premises. Second, a parsimonious conclusion is formulated based on information available within the mental model but not provided directly by the premises. Third, the conclusion is validated by checking that no alternative models of the premises violate this conclusion. If an alternative model of the premises refuting the conclusion is found the current conclusion is rejected, and phase two is repeated again (Johnson-Laird et al., 1992). Upon successful completion of the third phase, a fourth phase concerned with searching for an accepted theorem to support that no alternative model can refute the conclusion is required due to the nature of deductive mathematical arguments (Ufer et al., 2009). While the first three phases represent a heuristic approach to proving, the fourth phase represents a search for systematics theoretical arguments (i.e., analytical approach).

As stated by Ufer et al. (2009), geometry proof tasks are often accompanied by a geometric figure, hence generating a mental model during the first phase of deductive reasoning requires the integration of two types of information (i.e., premises): (a) verbal information (i.e., text), and (b) visual information (i.e., figure). Students then come up with intermediate conclusions in the second phase which are validated in the third phase. However, since the geometric figure is not merely a visual image, this framework suggested that mental models can be in the form of *figural concepts*, a term adopted from Fischbein (1993). According to Fischbein (1993), figural concepts are mental representations used in mathematical reasoning with geometric problems, consisting of a combination of conceptual and figural properties at the same time. Thus, mental models for geometry proofs can be conceptualized as flexible visual representations to the extent that they can be manipulated mentally but still preserve the properties given by the definitions and the premises of the relevant theorem (Fischbein, 1993).

The construction of geometry proofs can be performed through two different approaches: a *figure-based* approach based on mental images (phases 1-3) or an *analytical approach* based on propositional reasoning (phase 4). These two approaches are akin to the *semantic* and *syntactic* proof construction approaches advocated by Weber and Alcock (2004). Weber and Alcock (2004) stated that in syntactic proof construction, students manipulate only facts and definitions in a logical way. Whereas, in the semantic proof construction students 'instantiate' mathematical objects (i.e., informal mental representations of mathematical concepts) as a means to support their drawn conclusion (s) about the premise(s). Thus, unlike syntactic reasoners, semantic reasoners tend to use visual objects (i.e., geometric figure) during proof construction by manipulating these objects either mentally or on paper (Weber & Alcock, 2004).

3.1.4.1. Employing geometric proofs mental models during PF provision

When engaged in PF provision on peer solutions to geometry proofs, students normally validate a proof performed by a peer. Unlike proof construction, mathematical concepts and theorems are already available in proof validation, and students just need to judge their correctness (Selden & Selden, 2015a). Preservice mathematics teachers might employ the figure-based and analytical approaches (described earlier) during PF provision on peer solutions to geometry proofs. In the case of geometry proofs, students are also provided with a figure in combination with the proof. According to Koedinger and Anderson (1991) providing a visual model (the figure) to the student changes the way of testing the mental models. Findings of their study using intelligent tutoring system for geometry proof problem solving and interviews showed that students did not consider alternative visual models apart from the figure provided for them. Larkin and Simon (1987) assumed that making inferences from the figure (i.e., diagrammatic representation) in geometry proofs is easier for students than making inferences from statements (i.e., sentential representation). They attributed this case to the fact that related information is usually on the same or near locality in the figure, which allows students to view and process the information simultaneously. Accordingly, it seems reasonable to assume that preservice mathematics teachers are likely to focus on the figure during PF provision on a peer solution to a geometry proof. This assumption is in line with findings of studies showing that students rely mainly on empirical evidence (e.g., measuring the figure) when constructing geometry proofs (e.g., Klieme et al., 2003; Reiss et al., 2000) or performing construction tasks (e.g., Tapan & Arslan, 2009). However, in PF provision on peer solutions to geometry proofs, students are given some statements or theorems and a geometric figure. The quality of the peer solution might influence whether the PF provider focuses on the text component or the figure component of the proof. That is, when providing PF on an erroneous peer solution to geometry proof the PF providers are likely to experience cognitive conflict as a result of wrong statements or missing warrants. This inconsistency may require the PF providers to focus on the text and on the figure of the peer solution to be able to judge the correctness of the peer solution. Koedinger and Anderson (1991) stated that the figure contains some properties that might not be stated explicitly in the premises. Thus in the case of an erroneous peer solution, the PF provider might consult the figure to check the correctness of the statements in peer solution. On the other hand, providing PF on a correct or a near-correct peer solution to a geometry proof is unlikely to stimulate cognitive conflict, and the PF providers might adopt a figure-based approach while providing PF because as proposed by Larkin and Simon (1987) it is easier for students to make figure-based inferences. Nevertheless, it is unclear if the quality of the statements in the peer solution influences the approaches (figure-based vs. analytical) that preservice mathematics teachers might adopt while providing PF on a peer solution to a geometry proof. The assumptions regarding these two approaches that could be adopted by preservice teachers during PF provision should be empirically examined. The eye-tracking methodology is a good approach to test these theoretical assumptions.

3.1.5. Eye-tracking

Eye-movements have been increasingly used as a measure of overt attention while cognitively processing complex tasks (Rayner, 2009) because the rapidity of eye-movements matches the rapidity of cognitive processing (Just & Carpenter, 1976). As stated by Just and Carpenter's (1980) *eye-mind* assumption, a word is being processed as long as the eye remains fixated on it. Accordingly, time spent fixating on that word directly indicates the duration spent processing it. The eye-mind assumption postulates that the locus of the eye gaze is strictly tied to what is being cognitively processed (Just & Carpenter, 1976). When a student is engaged in a complex task such as reading or problem solving, then his/her attention and eye gaze are inseparable (Rayner, 2009).

Although eye-tracking methodology has been implemented mainly in research on reading and text comprehension (see Rayner, 1998; Rayner, 2009 for reviews), its use is becoming more popular in various domains including: mathematics problem solving (Beitlich, Obersteiner, & Reiss, 2015; Hegarty et al., 1992; Hegarty, Mayer, & Monk, 1995) and proof comprehension and validation (e.g., Hodds et al., 2014; Inglis & Alcock, 2012), science problem solving (e.g., Tsai et al., 2012), multimedia learning (e.g., Canham & Hegarty, 2010; Meyer et al., 2010), and more recently to infer mindful cognitive processing by students processing received PF (Bolzer et al., 2015). The proof validation and proof comprehension studies suggested that undergraduate students focus on surface features when validating proofs (Inglis & Alcock, 2012) and that short self-explanation training could lead students to process the proofs more deeply when reading for comprehension (Hodds et al., 2014). However, to date no study to our knowledge employed an eye-tracking methodology to examine PF providers' cognitive processing while providing PF on peer solutions to complex tasks like geometry proofs, and to explore whether specific types of PF are associated with gazing at specific components of geometry proofs (i.e., text or figure). There is a need to explore preservice mathematics teachers cognitive processing while providing PF on peer solutions to geometry proofs, to provide better support for them during PF activities and to develop better PF training for preservice mathematics teachers. Various eye-tracking measures can be used to infer cognitive processing, and they are described next.

3.1.5.1. Eye-tracking measures and terminology

The two main measures of eye-movement are *fixations* and *saccades*. Fixations are defined as time periods (200-300 milliseconds) during which the eye remains relatively still while information is being processed from the visual field (Rayner, 1998; Rayner, 2009).

Saccades are the rapid eye-movements, with high velocity (up to 500° per second), during which no processing takes place as the position of the eye-center (i.e., fovea) is shifted to a new location within the visual field (Rayner, 1998; Hyönä, 2010). The unit of analysis used to relate eye-movement measures to cognitive processing differs depending on the theory behind the analysis (Just & Carpenter, 1976). The unit of analysis could be a single word or a larger block that is called an *Area Of Interest (AOI)* (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, & De Weijer, 2011).

Different types of measures can be computed from the two main eye-movement measurements including, number of fixations, mean fixation duration, mean number of fixations, gaze duration, and total fixation time (Rayner, Chace, Slattery, & Ashby, 2006). Some of these terms (e.g., gaze duration vs. fixation duration) are used interchangeably due to their conceptual similarities (Holmqvist et al., 2011). Two famous measures are *fixation duration*, which is defined as the time duration during which an eye remains still, and *dwell time* (also known as gaze or glance duration) being defined as the time duration spent on an AOI from entry to exit of an AOI (Holmqvist et al., 2011). The term dwell time is also often confused with *total dwell time* which is the sum of all dwell times on a specific AOI accumulated over a trail (Holmqvist et al., 2011).

There is no agreement on the best measure to use for analyzing eye-tracking data in general because that depends on the research question and the size of AOI being used. Whereas fixation duration is very informative about the processing of a single-word, it might not be an optimal measure for larger AOIs that include words processed differently (Rayner, 1998). Furthermore, although total dwell time is a useful measure because it shows where students allocate their attention throughout the learning task, it does not provide information about moment-to-moment processing (Hyönä, 2010). When the unit of analysis is larger than a single word, total dwell time is usually used as a measure of eye-movement (Rayner et al., 2006). Due to the individual variation in most eye-tracking measures (Holmqvist et al., 2011), some studies use proportional measures in their investigations to account for this variation (e.g., Bednarik & Tukiainen, 2008; Hu, Wang, Fu, Quinn, & Lee, 2014; Yi, Liu, Li, Fan, Haung, & Gao, 2012).

3.1.5.2. Factors influencing eye-movements

Eye-movements differ depending on the nature of the encountered task, with complex tasks requiring more or longer fixations (e.g., silent vs. loud reading, or simple vs. complex problem solving; Rayner, 1998). This also applies to different parts of the same problem. For instance, the earlier context in a text influences how long readers look at later parts of the text (Rayner, 1998). Mathematics problem solving studies (e.g., Hegarty et al., 1992) also showed that inconsistent arithmetic problems resulted in longer fixations. Another factor that can influence how students distribute their attention on different components of the learning task is domain knowledge. Some mathematics and science problem solving studies showed that successful problem solvers focus more on the task-relevant information than the task-irrelevant information compared to unsuccessful problem solvers (e.g., Hegarty et al., 1995; Tsai et al., 2012), and that acquiring domain knowledge through instruction makes students focus more on task-relevant information and less on task-irrelevant information (Canham & Hegarty, 2010). Consequently, students' domain knowledge should be taken into account when eye-movements are being used as a measure of cognitive processing.

Despite the immediacy of eye-tracking measures, which makes them very essential to measure different cognitive processes (Just & Carpenter, 1976), this methodology has its limitations. Interpreting eye-tracking data requires a large degree of inference that must be backed up with clear and specific theoretical assumptions (Just & Carpenter, 1976) because this type of data does not provide any information about the success or failure of processing (Hyönä, 2010). For example, a student might spend a long time reading a proof without being able to understand it fully. This issue creates a need for complementary measures such as performance, or cued-retrospective reporting (Hyönä, 2010; Van Gog, Paas, & Van Merriënboer, 2005).

3.1.6. Current study

The aim of this study was to investigate how preservice mathematics teachers cognitively process and comprehend peer solutions to geometry proofs of different qualities while providing PF on one of them. The study also aims to further explore preservice mathematics teachers' individual characteristics including (a) experienced epistemic emotions curiosity, confusion, and anxiety, (b) PF providers' beliefs about PF provision, and (c) PF providers' perceptions of their PF. A between-subject design was implemented in which preservice mathematics teachers had to provide verbal PF on one of two fictional peer solutions to a geometry proof (a near-correct and an erroneous solution). To acquire control over the quality of peer solutions, which is required by the experimental design, fictional peer solutions were used. The fictional peer solutions were designed to have two parts: (a) a non-comparable (earlier) part which was near-correct for one condition but included errors in the other condition, and (b) a comparable (later) part which was exactly the same for both conditions. The eye-movements of participants were recorded and their comprehension of the geometry proof they provided PF on was measured. The following research questions were examined in this study:

3.1.6.1. Research questions and hypotheses

RQ 1. How does cognitive processing while providing PF (measured by proportional total dwell time) differ depending on the quality of a peer solution (near-correct vs. erroneous) to a geometry proof?

Hypothesis 1a: based on the findings that students spend longer time fixating on mathematical problems containing inconsistent information (Hegarty et al., 1992), as they engage in problem solving to resolve the inconsistency (Rayner, 1998), and the assumption that students might consult the geometric figure to check properties not mentioned specifically in the premises (Koedinger & Anderson, 1991) it was hypothesized that – for the non-comparable peer solution part – preservice mathematics teachers would have longer proportional total dwell time on the text, figure, and text and figure combined of the erroneous peer solution compared to the near-correct peer solution.

Hypothesis 1b: Based on the assumptions that the earlier task context influences the reading behavior of the later parts (Rayner, 1998), and that figure-based inferences are more efficient (Larkin & Simon, 1987) – it was hypothesized that preservice mathematics teachers who encountered more errors in the earlier non-comparable part of the peer solution would adopt an analytical approach when providing PF, and therefore have longer proportional total dwell time on the text component of the peer solution. Those who encountered no errors in the earlier non-comparable peer solution part would adopt a figure-based approach (Ufer et al., 2009), and instead will have longer proportional total dwell time on the figure corresponding to the comparable (later) part of the peer solution.

RQ 2. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the PF providers' comprehension of the proof?

Hypothesis 2: based on the findings that students have difficulties evaluating incorrect geometry proofs (Reiss et al., 2000; Zerr & Zerr, 2011), and the findings that students did not benefit more from studying erroneous examples compared to correct examples (Isotani et al., 2011), it was hypothesized that providing PF on the near-correct peer solution would result in better proof comprehension of the PF provider than providing PF on the erroneous peer solution.

RQ 3. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the content of the provided PF?

Hypothesis 3a: based on previous findings by Cho and Cho (2011) and Patchan et al. (2009), it was hypothesized that the type of PF provided on the erroneous peer solution would be different from PF provided on the near-correct peer solution. More specifically, it was hypothesized that the PF provided on the erroneous peer solution will contain more cognitive elaboration PF and self-efficacy PF and that PF provided on the near-correct peer solution will contain more cognitive surface PF and cognitive verification PF.

Hypothesis 3b: regarding the accuracy of PF, it was hypothesized based on previous findings (Reiss et al., 2000; Zerr & Zerr, 2011) that the PF provided on the near-correct peer solution will be more accurate than the PF provided on the erroneous peer solution, because it is easier for students to validate correct proofs.

RQ 4. What are the relationships between the PF providers' proportional total dwell time on the peer solution, PF providers' proof comprehension, and content of the provided PF?

Hypothesis 4a: based on the eye-mind assumption (Just & Carpenter, 1980), it was hypothesized that the proportional total dwell time on the text, figure, and text and figure combined of the peer solution would be associated with PF providers' proof comprehension for both conditions.

Hypothesis 4b: it was hypothesized that the type of PF (cognitive surface, cognitive verification, cognitive elaboration, self-efficacy, and affective) would be related to the proportional total dwell time on the text, figure, and text and figure combined of the peer solution.

Hypothesis 4c: it was hypothesized that accuracy of PF would be related to the proportional total dwell time on the text, figure, and text and figure combined of the peer solution for both conditions.

Hypothesis 4d: although to our knowledge no study so far has investigated the relationship between PF providers' understanding of the learning task and the type of PF they provide, the findings of Patchan and Schunn (2015) suggest that, for both conditions, positive correlations could be expected between PF providers' proof comprehension and cognitive elaboration PF and cognitive verification PF, as well as negative correlations between PF providers' proof comprehension and cognitive surface PF, self-efficacy PF and affective PF.

Hypothesis 4e: PF accuracy was expected to be positively associated with PF providers' proof comprehension because accurate PF requires a good understanding of the task.

RQ 5. What is the impact of the quality of the peer solution (near-correct vs. erroneous) to a geometry proof on the experience of the epistemic emotions curiosity and confusion?

Hypothesis 5: based on emotions research (D'Mello & Graesser, 2012; Kang et al., 2009), it was hypothesized that more curiosity and confusion would be experienced when PF is provided on the erroneous peer solution because students are more likely to experience cognitive conflict when evaluating the erroneous peer solution.

RQ 6. What are the relationships between the PF providers' epistemic emotions, proportional total dwell time on the peer solution, proof comprehension, beliefs about PF provision, perceptions of their PF, and the content of their PF?

Hypothesis 6a: based on the assertions by Muis et al. (2015), the proportional total dwell time on the text, figure, and text and figure combined of the peer solution was expected to correlate with curiosity and confusion positively.

Hypotheses 6b and 6c: Based on previous findings on mathematics problem solving (Muis et al., 2015), it was hypothesized that curiosity would be positively correlated with PF providers' proof comprehension (hypothesis 6b), and confusion would be negatively correlated with PF providers' proof comprehension (hypothesis 6c).

Hypotheses 6d and 6e: regarding the relationship between the epistemic emotions curiosity, confusion, and PF content (including type and accuracy), it was hypothesized that PF content would be related to curiosity (hypothesis 6d) and confusion (hypothesis 6e), because these emotions predicted different problem solving strategies (Muis et al., 2015).

Hypothesis 6f: The relationship between PF content and anxiety was not tested previously to our knowledge, yet since anxious students are expected to express concerns about their ability to construct the proof or to provide PF on it, anxiety was expected to correlate negatively with cognitive elaboration PF and PF accuracy, and positively with self-efficacy PF.

Hypotheses 6g, 6h, and 6i: Regarding the relationship between beliefs about PF provision, PF providers' perceptions of their PF, and PF content (including type and accuracy), it was expected that the beliefs about learning from PF provision, confidence regarding PF provision and engaging in reasoning during PF provision would correlate positively with cognitive verification PF and cognitive elaboration PF, and negatively with self-efficacy PF (hypothesis 6g). It was also hypothesized that beliefs about PF provision would correlate positively with PF accuracy (hypothesis 6h) and that PF providers' perceptions of their PF would correlate positively with PF accuracy and beliefs about having confidence regarding PF (hypothesis 6i), because these types of beliefs and perceptions are likely to be associated with performance.

3.2. Method

3.2.1. Participants

Participants were preservice mathematics teachers for high school (Realschule or Gymnasium) from a large university in southern Germany. Participation in the study was voluntary and participants were compensated with €15 for their participation. A total of 53 participants (14 males, 39 females, mean age 24.40, age range: 18-49 years, *SD* = 5.33) were included in the analyses. All participants signed an informed consent before taking part in the experiment.

3.2.2. Design

The study adopted a between-subject experimental design with two conditions: condition A (near-correct fictional peer solution to a geometry proof) and condition B (erroneous fictional peer solution to a geometry proof). Participants were randomly assigned to one of the two conditions: condition A (*n* = 27), and condition B (*n* = 26). The experiment consisted of two phases: an on-screen phase in which we used eye-tracking methodology to measure participants' cognitive processing of a peer solution while providing verbal PF on that solution; and a paper-and-pen phase, in which their comprehension of the proof they provided PF on, their basic geometric knowledge, their beliefs about PF provision, their perceptions of their PF, and their epistemic emotions were measured. Initially, one additional participant was assigned condition B, but had to be excluded because large parts of the screen were not visible in the video recordings which affected the accuracy of the eye-tracking measures.

3.2.3. Materials

3.2.3.1. Experimental tasks: peer solutions to a geometry proof

The experimental tasks were two fictional peer solutions to a geometry proof as shown in Figure 5. The experimental tasks were designed by an expert in the domain of mathematics.

In condition A the fictional peer solution consisted of nine steps, while the fictional peer solution in condition B consisted of eight steps (i.e., one missing step). The solutions were designed in such a way that both conditions have a non-comparable part followed by a comparable part. In condition A, the non-comparable part (steps 1-3) had very few weak warrants which are considered insignificant errors according to the sociomathematical norms in the participants' teacher-training program. Conversely, in condition B the non-comparable part (steps 1-2) had several unwarranted claims and one missing step. The comparable part (constituent steps: 4-9 in condition A; 3-8 in condition B) was exactly the same in both conditions and was almost correct with only minor notation errors in both solutions.

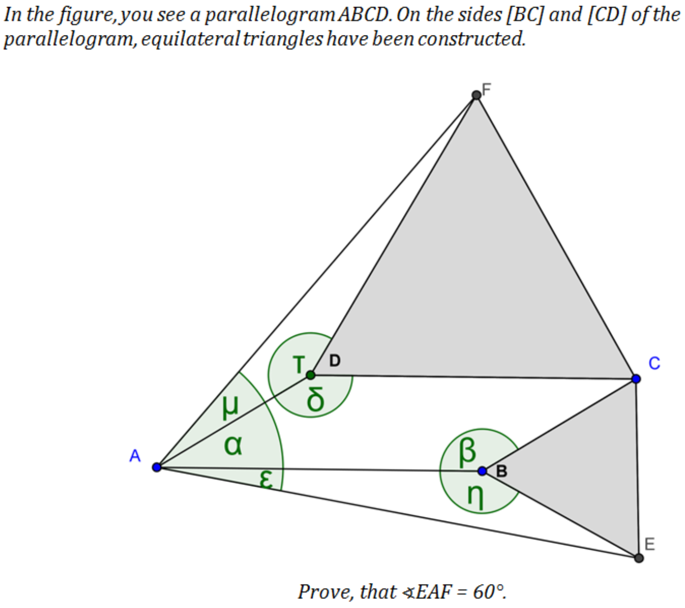

Figure 5. The geometry proof for which the two experimental fictional peer solutions were created.

3.2.3.2. Peer feedback instruction and example task

To control for participants' prior experience with PF, every participant received an introduction to PF – including a statement about the importance of PF provision to proof learning, as well as definition of constructive PF (i.e., a feedback that explains what was correct and why, always includes justifications for statements, is supportive and helps students to improve their solution). As part of the instruction, the participants received an example task consisting of a fictional peer solution to a geometry proof different from the experimental task. This peer solution was also created by the same domain expert who created the peer solutions for the experimental conditions. The purpose of using this fictional peer solution was to illustrate some examples of constructive and unconstructive PF. The PF examples were reviewed by two mathematics experts to ensure their accuracy.

3.2.3.3. Manipulation checks

Manipulation checks were performed in the second phase of the experiment. Each participant was asked to rate the quality of peer solution they provided PF on, by answering one item: "Please rate the performance of your peer from 0 to 100 by crossing the appropriate place on the line" on 10-cm visual analog (VAS) ranging from 0 to 100. Additionally, the level of experienced cognitive conflict was measured by one 10-cm VAS item: "How similar to this solution would you solve this task?" ranging from "very different" to very "similar".

3.2.4. Measures

3.2.4.1. Proportional total dwell time measured by eye-tracking

Participants' eye-movements were recorded using a monocular head-mounted Dikablis 50 Hz eye-tracker (by Ergoneers LLC) with 3-mm objective. The eye-tracker operated at the default speed (25 Hz) set by the company for the recording software version 2.0. Total dwell time, in seconds, was computed for predefined areas of interest (AOIs) for each condition. The figural components were the same for both experimental conditions and did not change in the non-comparable and the comparable peer solution parts. The only thing that changed was the text component. However, since the participants would look at the figure in relation to the proof's text, separate measures were calculated for the figure corresponding to each part of the peer solution. For both conditions, total dwell time was computed for the following AOIs: (1) non-comparable peer solution part (text and figure combined), (2) comparable peer solution part (text and figure combined), (3) the figure corresponding to the non-comparable peer solution part, (4) the text of the non-comparable peer solution part, (5) the figure corresponding to the comparable peer solution part, and (6) the text of the comparable peer solution part.

To account for individual differences in reading time (Holmqvist et al., 2011), *proportional total dwell time* (PTDT) for each AOI was calculated by dividing the total dwell time on a specific AOI by the overall dwell time on all AOIs in the relevant parts of the peer solution. More specifically, the total dwell time on (text and figure combined) was divided by overall dwell time on the entire peer solution, whereas the total dwell time on text or on figure of comparable and non-comparable parts of the peer solution was divided by the total dwell time on that part of the peer solution (i.e., total dwell time on text and figure combined).

Since the near-correct peer solution in condition A had one more step than the erroneous peer solution in condition B in the non-comparable peer solution part, the size of the text-AOI for condition A was larger than that for condition B. There are two ways to deal with this issue, either by dividing the PTDT by the number of characters, or by the size of the AOI. Dividing by the number of characters is not recommended because it can produce a non-linear relationship with reading time (see Rayner, 1998; Trueswell, Tanenhaus, & Garnsey, 1994). Accordingly, the PTDT of the non-comparable part of the peer solution for both conditions was divided by the AOI's area in pixels, to control for differences produced by different sizes of AOIs in that part of peer solution. This was only done to the non-comparable part of the peer solution because the comparable part of the peer solution was identical in both conditions.

3.2.4.2. Proof comprehension

Participants' understanding of the geometry proof used in the experiment was measured with a proof comprehension test. The test was constructed for the current study and was mainly based on some elements of Yang and Lin's (2008) model of reading comprehension of geometry proof. More specifically, items in the test addressed the following three dimensions of their model: (1) *surface elements*: understanding meanings of different components of proof including terms, symbols and figures, or calculations of geometric measurements, (2) *recognizing elements*: understanding of different types of proof statements including premises, properties or conclusions, and (3) *chaining elements*: understanding the logical connection between different elements of the proof, and viewing the figure as a reference object (Yang & Lin, 2008). Sixteen open-ended questions were created by an expert in mathematics. The questions were assigned different weights depending on what the question asked for. One or two points were awarded for each correct expected assertion, and 0 was awarded for missed or wrong assertions. Questions 1, 2, and 6 had a maximum of two points each; questions 3 and 4 had a maximum of six points each; and questions 5 and 7-16 had a maximum of four points each. The test consisted of two questions about the non-comparable peer solution part (questions 2 and 3), six questions about the comparable peer solution part (questions 4-9), and eight questions testing the general understanding of the proof and the underlying theorems (question 1 and questions 10-16). Two independent coders (student assistants) scored 10% of the tests reaching a good inter-rater agreement (Krippendorff's $\alpha = .80$). Afterwards, one of the student assistants scored the tests of all participants.

A principal component analysis (PCA) was conducted using data from the current study to ensure that a total sum score can be used to represent participants' understanding of the proof as assessed by the proof comprehension test. Parallel analysis was conducted following O'Connor's (2000) procedure. By comparing the raw data eigenvalues to the percentile values produced by the parallel analysis, a one-factor solution seemed to best explain the data (raw data eigenvalue > percentile value). As suggested by the parallel analysis, a PCA with the one-factor solution was computed. The items were factorable (KMO measure of sampling adequacy = .89, Bartlet's test of sphericity was significant p =.000) and had a positive determinant of the correlation matrix (.005). Most items had factor loadings > .40 and a value of at least .50 for the diagonal elements of the anti-image correlation matrix (Field, 2009). Questions 1, 10 and 11 had low factor loadings, very low communalities and an anti-image correlation value < .50. The internal consistency of the test was determined using Cronbach's α. Questions 1, 10 and 11 had corrected item-total correlation < .30 and removing them resulted in an improved value of Cronbach's α from .77 to .82. Based on the results of the PCA and the reliability test, questions 1, 10 and 11 were excluded. A total sum score of the proof comprehension test was computed by summing up the total of the remaining items with a maximum score of 52 points.

3.2.4.3. Basic geometric knowledge

A basic geometric knowledge test was used to measure students' basic geometric knowledge. This test was used in previous research (e.g., Ufer et al., 2008). The test consists of 49 true/false items which were scored dichotomously as 0 if answered incorrectly and 1 if answered correctly (Cronbach's α = .79). The test measured different basic geometric knowledge including properties of triangles, properties of a parallelogram, transversals, and quadrangles. A total sum score of the 49 items was computed as a measure of participants' basic geometric knowledge. This measure was used as a covariate in the analyses.

3.2.4.4. Peer feedback content: type and accuracy

The content of PF was determined regarding (a) the type of the provided PF, and (b) the accuracy of the provided PF.

Peer feedback type

The PF type was operationalized as having two dimensions – *purpose* and *style* of PF – based on a coding scheme to analyze PF messages developed by Strijbos et al. (2012). The coding scheme, which was created for bachelor's thesis proposal, was adapted for use with PF on peer solutions to geometry proofs. According to the coding scheme, the purpose of PF could have (a) a *cognitive focus* on the content knowledge including knowledge about the overall proof or parts of the proof, (b) a *metacognitive focus* on the general learning strategies related to the learning task as well as monitoring and evaluating the learning process, (c) an *affective focus* on motivating and encouraging the (fictional) peer for future performance, and (d) a *self-efficacy focus* on the providers' confidence in their ability to provide PF or to solve the proof. Cognitive and metacognitive PF can have two styles: *verification* or *elaboration. Verification* represents confirming or disconfirming statements about the correctness or incorrectness of the entire peer solution or parts of the peer solution, or the learning approaches used to deal with the task. *Elaboration* represents more detailed comments about parts of the peer solution that follow verifications or other elaborations in the form of correction, confirmation, justification, questioning, or suggestions. PF comments can also include general statements that are neither verifications nor elaborations, or comments on the surface features of the peer solution (e.g., presentation style, spelling or grammar mistakes, etc.). The general statements normally serve as an anchor to a subsequent verification or elaboration (Strijbos et al., 2012). The general statements were not included in statistical analyses of PF in the current study because they do not include any evaluative information about the peer solution.

Coding procedure

Two student assistants transcribed the verbal PF. The transcripts were then segmented following the procedure by Strijbos et al. (2006), with the smallest meaningful segment as the unit of analysis. Two independent coders (student assistants) segmented parts of the data over several independent rounds, with 10 % of the data each time, until an acceptable percentage agreement was reached (92.5 % lower bound; 93 % upper bound). Afterwards, one of the student assistants segmented the remaining data.

The segmented PF data was subsequently coded using the coding scheme adapted from Strijbos et al. (2012). Two independent coders (student assistants) segmented part of the data over several rounds, with 10% of the data each time, until an acceptable inter-rater reliability was reached (Krippendorf's α = .81). No metacognitive PF was observed in the current data. The reliability of the coding was calculated for different levels of the coding scheme: main categories (i.e., cognitive, affective, and self-efficacy; Krippendorf's α = .68), sub-categories (e.g., cognitive verification, cognitive elaboration, general cognitive statement, etc.; Krippendorf's α = .81), and lowest level of sub-categories (e.g., cognitive confirmation, cognitive justification, content self-efficacy, PF self-efficacy, etc.; Krippendorf's α = .76).

Although the main categories typically have higher reliability values than the smaller categories, because they are easier to distinguish, this was not the case for the current sample. The majority of the PF belonged to the main category 'cognitive', which might have resulted in the very few disagreements on the other main categories (self-efficacy and affective) causing a severe correction of the Krippendorff's alpha value. The chance-correction level is higher for three categories (compared to the five subcategories within the cognitive category). Thus, it might be the case that the data did not have enough cases for the other main categories for a more accurate estimate of the inter-rater reliability. After reaching an acceptable inter-rater reliability value for the PF levels used for analysis in this study (Krippendorff's α = .81), the same student assistants coded half of the remaining data each.

Peer feedback accuracy

For PF accuracy we adopt the definition of PF quality from Van Steendam et al. (2010), but limiting PF accuracy to the number of detected errors or correct statements. Accordingly, the provided PF was assessed in terms of the number of accurate comments about the correct, incorrect or missing aspects in the peer solution that the participants provided PF on. The same student assistants who coded for PF type also coded the accuracy of the comments. Over several rounds, each time 10% of the data was coded until an acceptable level of inter-rater reliability was reached (Krippendorf's $\alpha = .76$). Afterwards, each student assistant coded half of the remaining data.

The PF accuracy score was not directly comparable because the two peer solutions had comparable and non-comparable parts. The peer solution in condition A was longer, and the other peer solution in condition B had more errors. The maximum PF accuracy score for the peer solution in condition A was 14 points, whereas it was16 points for condition B. To compare the two conditions, proportional PF accuracy scores were computed by dividing the number of accurate comments made by the maximum number of possible correct comments.

3.2.4.5. Beliefs and perceptions about peer feedback provision

Since students' beliefs represent a trait disposition that is relatively stable across activities, and their perceptions represent a state disposition that is instant and connected to task-specific experiences (Smith, 2001), both were measured in the current study because they convey different information about the students' perspective on the PF provision activity.

Beliefs about peer feedback provision

Participants' beliefs about PF provision were measured using the PFPQ developed and tested for Study 1. The questionnaire consists of three subscales: learning from peer feedback provision (LPF), confidence regarding peer feedback provision (CPF), and engaging in reasoning during peer feedback provision (RPF). Part of the CPF scale was adapted from the Feedback Orientation Scale (Linderbaum & Levy, 2010), and the other two scales were created for Study 1. Although parallel analysis (O'Connor, 2000) with the data from this study revealed a three-factor solution, a PCA was not conducted for the current sample as the sample size was small relative to the number of items. Instead, the same scales as composed from the pilot in

Study 1 were used in this study, which rendered equally high internal consistencies. The LPF scale consisted of 10 items (e.g., “I want to change my own work after providing peer feedback”; Cronbach’s α = .80); the CPF scale consisted of 17 items (e.g., “I believe that I have the competence to provide clear peer feedback”; Cronbach’s α = .91), and the RPF scale consisted of 13 items (e.g., “Providing peer feedback helps me to be critical about my own arguments”, Cronbach’s α = .76). The means for the three scales were computed and used for further analyses.

Peer feedback providers’ perceptions of their peer feedback: quality, accuracy, and usefulness

Participants were asked to rate the PF they provided in terms of its quality, accuracy and usefulness using three items: “How confident are you about the quality of the feedback you provided?”, “How confident are you that your feedback is correct?”, and “How useful do you think your feedback will be for this student?”. The items were scored on a 10-cm visual analogue scale (VAS) ranging from “not confident at all” to “very confident”.

3.2.4.6. Epistemic emotions

Epistemic emotions experienced while providing PF were measured with the short version of Pekrun and Meier’s (2011) epistemic emotions scale. This scale consists of seven discrete emotions each measured by a single-word item (curious, surprised, confused, frustrated, excited, anxious, and bored) and rated in terms of intensity on a 5-point Likert scale ranging from *not at all* (1) to *very strong* (5).

3.2.5. Procedure

The study consisted of two phases during which participants were tested individually in a quiet room. In phase 1 (on-screen), preservice mathematics teachers were presented with an example of a peer solution to a geometry proof followed by written examples of constructive and unconstructive PF on that peer solution. Afterwards, the participants were presented with one fictional peer solution – a near-correct peer solution or an erroneous peer solution – and were asked to provide verbal PF on that peer solution. The participants could navigate freely through different slides in phase 1 using the space bar. Before the start of phase 1, the eye-tracker was calibrated for maximum accuracy. The participants' voices were recorded during the PF provision phase using an external voice recorder. The eye-tracker was removed at the end of phase 1. Phase 2 (paper-and-pen) consisted of answering the proof comprehension, the basic geometric knowledge tests, the PF provision beliefs, the PF providers' perceptions of their PF, and the epistemic emotions questionnaires. In phase 2, the participants received a printed copy of the peer solution they had provided PF on in phase 1. Participants completed the two phases at their own pace and interaction with the researcher was kept to a minimum. The entire experiment lasted between 90 to 120 minutes. The experiment was piloted with eight preservice mathematics teachers before the main data collection to ensure the accuracy of the eye-tracking measures and the comprehensibility of the materials used. These students were not included in the sample of this study.

3.2.6. Analyses

Although some of the statistical tests used in this study (e.g., ANCOVA) are frequently reported to be robust to the violation of some of their assumptions, in cases of unequal sample sizes or when the difference between groups' variances is drastically large Type I or Type II errors are inflated (Field, 2009; Howell, 2008). Therefore, whenever one of the assumptions of parametric statistical tests (e.g., normality or homogeneity of variance) was violated by one of the variables under investigation, non-parametric alternatives such as the Wilcox robust ANCOVA analysis (Wilcox, 2012; Wilcox, 2013) or Spearman's rank-order correlations were used.

The Wilcox robust ANCOVA uses quantile regression instead of the least square correlation used in conventional ANCOVA, and it does not assume homoscedasticity or homogeneity of slopes regression (Wilcox, 2013). This method is recommended to be used when the underlying assumptions of ANCOVA are violated – instead of data transformation which does not yield accurate results most of the time – because robust ANCOVA does not assume linear regression (Field, 2009; Wilcox, 2012). According to Wilcox (2012), the robust ANCOVA uses a smoothing function that compares the 20% trimmed means at different design points (i.e., different values of the covariate) where the regression lines of both groups are comparable. It is recommended that the sample size for each group for the selected design points is not less than 12 (Mair & Wilcox, 2015). Simulations showed that this procedure is robust with non-linear and heteroscedastic data; even with relatively small sample sizes (as small as n = 20), or heavy-tailed distributions (Wilcox, 2013). The downside of the Wilcox robust ANCOVA is that it still has no measure of effect size.

A robust tests R package WRS2 version 0.3-2 (Mair & Wilcox, 2015) was used to run the robust ANCOVA tests. Four design points were used for the analyses in the current study, which were automatically specified by the function with a minimum of 12 participants in each group around the design points (for details see Wilcox, 2012). The p-value is not adjusted for multiple tests at the different design points, but the confidence intervals are adjusted (Mair & Wilcox, 2015). Therefore, p-values produced by this test will be interpreted in consultation with the confidence intervals. The R Project for Statistical Computing (version 3.3.1) was used to run the robust ANCOVA analyses, whereas the other analyses were conducted with IBM SPSS Statistics 23.

To acquire more information about the strength of evidence supporting the tested hypotheses, Bayes factors (BF_{10}) are reported for each tested model in comparison to the null hypothesis. The Bayes factors are interpreted using Jeffreys' (1961) classifications of evidence and were calculated using JASP (0.8.0.0). Sine Bayesian analyses can also be influenced by violations of assumptions including normality or homogeneity of variance; Bayes factors were only calculated whenever the assumptions were met.

Although omega-squared (ω^2) is a less biased measure – compared to eta-squared – especially with small sample sizes (Okada, 2013), it cannot be used when the number of participants in each group is not equal (Field, 2009). Following the recommendations by Lakens (2013) partial eta squared (η_p^2) is used as a measure of effect size, with values of .01, .06, and .14 corresponding to small, medium, and large based on Cohen's (1988) benchmark. Participants' basic geometric knowledge was included as a covariate in the analysis of PTDT, proof comprehension, PF content, and epistemic emotions curiosity and confusion.

3.3. Results

The purpose of this study was to investigate the impact of peer solution quality (near-correct vs. erroneous proof) on PF providers' cognitive processing and comprehension of the learning task during PF provision. We also investigated the role of PF providers' individual characteristics including their beliefs about PF provision, their perceptions of their PF, and their experienced epistemic emotions curiosity, confusion and anxiety in the PF provision activity.

3.3.1. Data inspection

The standardized skewness and kurtosis values were determined for each research condition separately and were within the acceptable range (± 3; Tabachnick & Fidell, 2013) for the proof comprehension total score, the PF accuracy score, the PFPQ subscales, the epistemic emotions curiosity, confusion and anxiety, the estimated peer performance item, the cognitive conflict item, and PF providers' perceptions of their PF including quality, accuracy, and usefulness. The eye-tracking measures (i.e., PTDT) had similarly acceptable standardized skewness and kurtosis values, apart from the PTDT spent on text (standardized skewness =

3.20) and on figure (standardized skewness = -3.20) for the non-comparable peer solution part in condition A. The standardized skewness and kurtosis values were outside the acceptable range for the basic geometric knowledge (standardized skewness = -4.00, standardized kurtosis = 4.50) for condition A, and for PF types including cognitive surface PF (condition A: standardized skewness = 3.60; condition B: standardized skewness = 5.02; standardized kurtosis = 6.61), affective PF (condition A: standardized skewness = 7.59; standardized kurtosis = 5.50), and self-efficacy PF (condition A: standardized skewness = 7.36; standardized kurtosis = 14.63; condition B: standardized skewness = 4.48). Three extreme univariate outliers ($z > | 3.29 |$; Field, 2009) were identified; one for basic geometric knowledge, one for cognitive surface PF, and one for affective PF. The values of the outliers were checked to ensure that they were not caused by an error in data-entry. The outlier identified for basic geometric knowledge was adapted to the second lowest value (Field, 2009) as a result of which standardized skewness and kurtosis became within the ± 3 range. Adapting the values of the other two outliers did not improve the standardized skewness and kurtosis, so it was decided to retain the original values. Non-parametric tests were used to analyze the variables with the extreme outliers (i.e., cognitive surface PF and affective PF). No multivariate outliers were identified. Participants in condition B did not use affective PF; therefore, no group comparisons were possible for this type of PF.

3.3.2. Manipulation checks

To ensure that participants in each condition did not differ regarding their basic geometric knowledge, an independent samples t-test was performed, with basic geometric knowledge as the dependent variable and experimental condition as the independent variable. The assumption of equality of variances was met according to Levene’s test of equality of variances (p = .882). There was no significant difference in basic geometric knowledge between participants in condition A (M = 42.26, SD = 3.99) and participants in condition B (M = 41.46, SD = 3.71), $t(51) = 0.75$, $p = .455$, d = 0.125.

To investigate whether participants were able to identify the quality of the peer solution presented to them (i.e., near-correct vs. erroneous), an independent samples t-test was performed with the estimated performance for the fictional peer as the dependent variable and the quality of peer solution as the independent variable. According to Levene's test of equality of variances, the assumption of equality of variances was met ($p = .290$). There was no significant difference in the estimated performance for the fictional peer solution between condition A peer (M = 67.89, SD = 15.66) and condition B (M = 64.79, SD = 14.14), $t(51) = 0.76$, $p = .454$, d = 0.207.

To test whether the erroneous peer solution stimulated more cognitive conflict than the near-correct peer solution, an independent samples t-test was performed, with cognitive conflict as the dependent variable and quality of peer solution as the independent variable. The assumption of equality of variances was met according to Levene's test of equality of variances ($p = .289$). A significant difference was found in cognitive conflict experienced by participants in each experimental condition, $t(51) = 2.21$, $p = .031$, d = .608. Participants who provided PF on the erroneous peer solution (M = 4.13, SD = 2.14) experienced significantly more cognitive conflict compared to those who provided PF on the near-correct peer solution M = 5.33, SD = 1.77).

3.3.3. RQ 1: The impact of peer solution's quality on reading behavior measured by proportional total dwell time (PTDT)

Regarding the first research question, it was hypothesized that compared to participants in condition A, those in condition B would have longer PTDT on text, figure, and on text and figure combined of the (earlier) non-comparable peer solution part (hypothesis 1a). It was, also, hypothesized that participants who encountered more errors in the earlier non-comparable peer solution part (i.e., condition B) would adopt an analytical approach while providing PF on the (later) comparable peer solution part and would, therefore, have longer PTDT on the text of the comparable peer solution part. On the other hand, participants who encountered minor errors in the (earlier) non-comparable peer solution part (i.e., condition A) would adopt a figure-based approach while providing PF on the comparable peer solution part, and, therefore, have longer PTDT on the figure of the comparable peer solution part (hypothesis 1b).

Differences in PTDT on non-comparable peer solution part

One-way ANCOVA tests were conducted for the PTDT on text, figure, and on text and figure combined of the non-comparable peer solution part, with basic geometric knowledge as a covariate. The assumption of homogeneity of regression slopes was met for all variables (p > .05). Wilcox robust ANCOVA tests were used for the three dependent variables, because the assumption of homogeneity of variance was violated ($p < .01$). The four comparison design points used to compare the 20% trimmed means are 42, 43, 44, and 45. Significant effects of the quality of the peer solution (near-correct vs. erroneous) were found after controlling for basic geometric knowledge at the four comparison points (i.e., values of basic geometric knowledge) on the PTDT on text and figure combined, $F_{42} = 3.94$, $p = .001$, 95% CIs [-4e^{-04}, -1e^{-04}]; $F_{43} = 4.20$, $p = .002$, 95% CIs [-4e^{-04}, -1e^{-04}]; $F_{44} = 4.09$, $p = .001$, 95% CIs [-3e^{-04}, -1e^{-04}]; $F_{45} = 4.61$, $p = .001$, 95% CIs [-3e^{-04}, -1e^{-04}]. Participants in condition B ($M = .00043$, $SD = .0001$) had significantly longer PTDT on text and figure combined of the non-comparable peer solution part compared to participants in condition A ($M = .00026$, $SD = .0001$) (Figure 6).

Similarly, significant effects of the quality of the peer solution was found on the PTDT on the text of the non-comparable peer solution part at the four comparison points, after controlling for basic geometric knowledge, $F_{42} = 6.43$, $p = .000$, 95% CIs [-.001, -4e^{-04}]; F_{43} =

7.12, p = .000, 95% CIs [-.001, -4e^{-04}]; F_{44} = 7.29, p = .000, 95% CIs [-.001, -5e^{-04}]; F_{45} = 5.91, p = .001, 95% CIs [-.001, -4e^{-04}]. Participants in condition B (M = .0011, SD = .0003) had significantly longer PTDT on the text of the non-comparable peer solution part than those in condition A (M = .0004, SD = .0001) (Figure 7). Significant effects of the quality of the peer solution were also found on the PTDT on the figure of the non-comparable peer solution part at the four comparison points on, after controlling for basic geometric knowledge, F_{42} = 13.18, p = .000, 95% CIs [-.002, -9e^{-04}]; F_{43} = 13.46, p = .000, 95% CIs [-.001, -1e^{-04}]; F_{44} = 16.72, p = .000, 95% CIs [-.001, -1e^{-03}]; F_{45} = 13.51, p = .000, 95% CIs [-.002, -9e^{-04}]. Participants in condition B (M = .002, SD = .0003) had significantly longer PTDT on the figure of the non-comparable peer solution part compared to participants in condition A (M = .001, SD = .0001) (Figure 7).

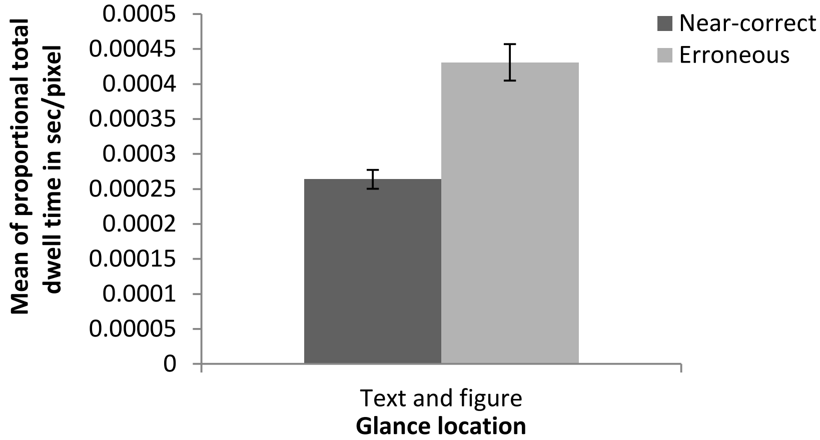

Figure 6. The means of proportional total dwell time spent looking at text and figure combined of the non-comparable peer solution part for condition A and condition B in seconds/pixel. Error bars are standard errors of means.

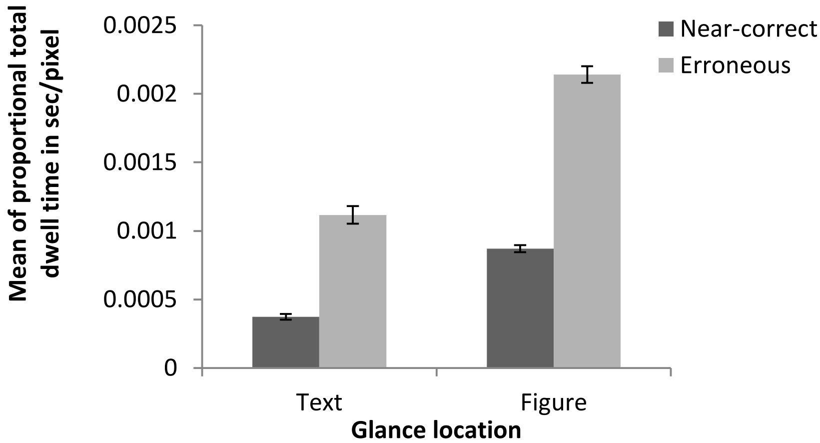

Figure 7. The means of proportional total dwell time spent looking at the text and the figure of the non-comparable peer solution part for condition A and condition B in seconds/pixel. Error bars are standard errors of means.

Differences in PTDT on comparable peer solution part

One-way ANCOVA tests were conducted for the PTDT on text, figure and text and figure combined of the comparable peer solution part, with basic geometric knowledge as a covariate. Levene's test of equality of variances indicated that the assumption of homogeneity of variance was met for the PTDT on text (p = .636), PTDT on figure (p = .636), and PTDT on text and figure combined (p = .180). The assumption of homogeneity of regression slopes was also met for all variables (p > .05). A significant effect for peer solution quality (near-correct vs. erroneous) was found after controlling for basic geometric knowledge on the PTDT on text and figure combined of the comparable peer solution part, F(1, 50) = 25.51, p =.000, η_p^2 = .338, B_{10} = 814.48 (decisive evidence for H_1). Participants in condition B (M = .70, SD = .09) had significantly longer PTDT on text and figure combined of the peer solution comparable part compared to participants in condition A (M = .56, SD = .12) (Figure 8). The covariate basic geometric knowledge had no impact on PTDT for text and figure combined, F(1, 50) = .37, p = .546, η_p^2 = .007, B_{10} = 0.236 (moderate evidence for H_0).

A significant effect for the peer solution quality (near-correct vs. erroneous) was found on the PTDT on the text of the comparable peer solution part, after controlling for basic geometric knowledge, F(1, 50) = 105.51, p = .000, η_p^2 =.678, B_{10} = 4.69e^{+10} (decisive evidence for H_1). Participants in condition B (M = .68, SD = .08) had significantly longer PTDT on the text of the comparable peer solution part compared to participants in condition A (M = .43, SD = .10) (Figure 9). The covariate basic geometric knowledge had no impact on the PTDT on the text of the comparable peer solution part, F(1, 50) = .23, p = .631, η_p^2= .005, B_{10} = 0.385 (anecdotal evidence for H_0). Similarly, a significant effect for the peer solution quality (near-correct vs. erroneous) was found on the PTDT on the figure of the comparable peer solution part after controlling for basic geometric knowledge, F(1, 50) = 105.51, p = .000, η_p^2 = .678, B_{10} = 4.90e^{+10} (decisive evidence for H_1). Participants in condition B (M = .32, SD = .08) had significantly shorter PTDT on the figure of the comparable peer solution part compared to participants in condition A (M = .56, SD = .10) (Figure 9). The covariate basic geometric knowledge had no impact on the PTDT on the figure of the comparable peer solution part, $F(1, 50) = .23$, $p = .631$, $\eta_p^2 = .005$, $B_{10} = 0.385$ (anecdotal evidence for H_0).

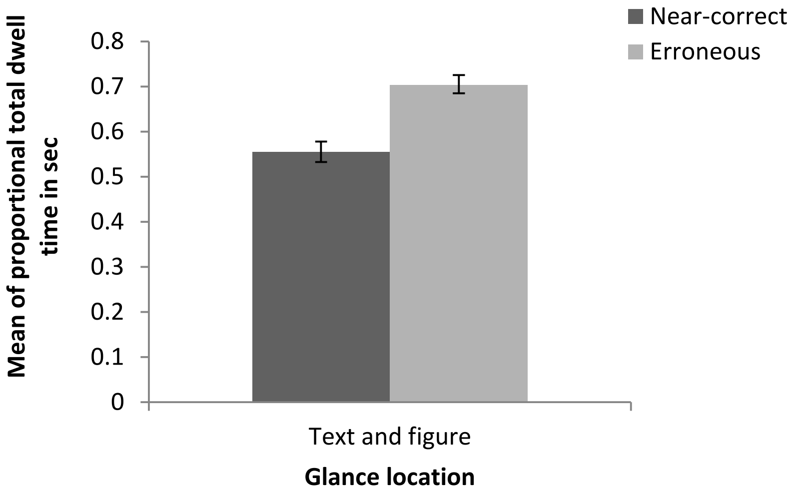

Figure 8. The means of proportional total dwell time spent looking at text and figure combined of the comparable peer solution part for condition A and condition B in seconds. Error bars are standard errors of means.

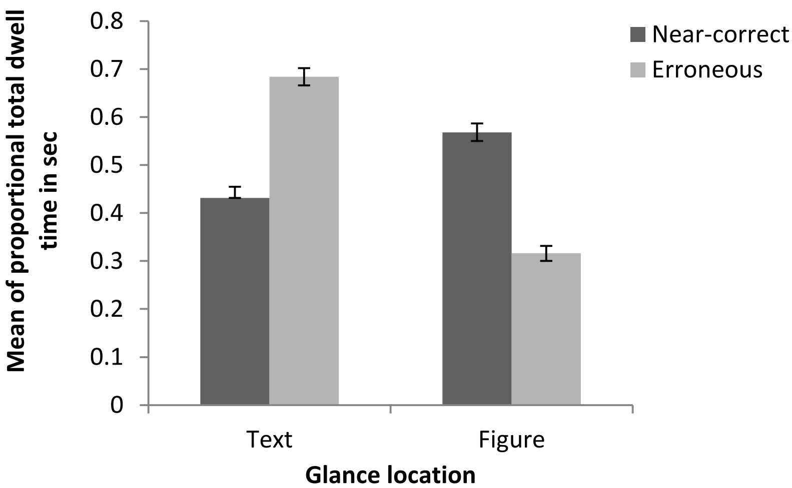

Figure 9. The means of proportional total dwell time spent looking at the text and the figure of the comparable peer solution part for condition A and condition B in seconds. Error bars are standard errors of means.

3.3.4. RQ 2: The impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the PF providers' comprehension of the proof

To investigate whether participants in condition A had better proof comprehension than participants in condition B (hypothesis 2) a One-way ANCOVA was conducted with peer solution quality as the independent variable, PF providers' proof comprehension as the dependent variable, and basic geometric knowledge as a covariate. Levene's test indicated homogeneity of variances (p = .467) and the assumption of homogeneity of regression slopes was met (p = .936). There was a significant effect of peer solution quality on proof comprehension, after controlling for basic geometric knowledge, $F(1,50) = 4.99$, $p = .030$, $\eta_p^2 = .091$, $B_{10} = 28487.25$ (decisive evidence for H_1). As illustrated in Figure 10, participants who provided PF on the near-correct peer solution (M = 30.22, SD = 11.10) performed significantly better in the proof comprehension test than those who provided PF on the erroneous peer solution (M = 23.65, SD = 10.11). The covariate basic geometric knowledge was significantly related to PF providers' proof comprehension, $F(1, 50) = 30.44$, $p = .000$, $\eta_p^2 = .378$, $B_{10} = 13638.01$ (decisive evidence for H_1).

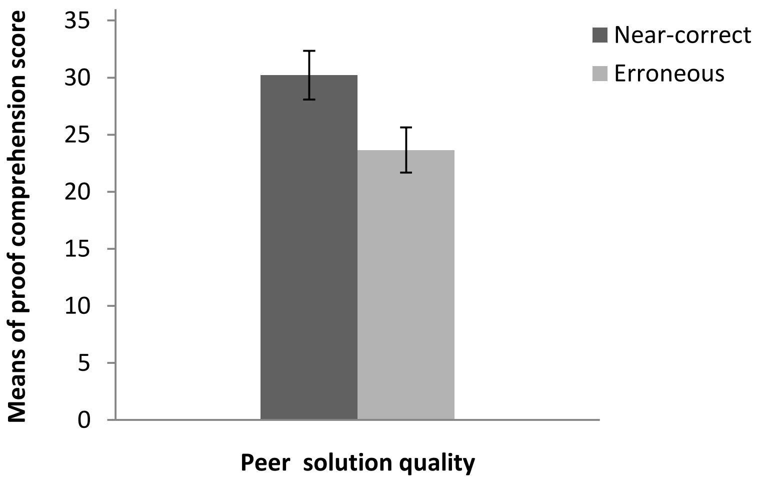

Figure 10. The means of proof comprehension scores of participants who provided PF in condition A and condition B. Error bars are standard errors of means.

3.3.5. RQ 3: The impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the content of the provided PF

Regarding the impact of peer solution quality on the content of provided PF (i.e., PF type and accuracy), it was hypothesized that participants in condition B would provide more cognitive elaboration PF and self-efficacy PF, whereas participants in condition A would provide more cognitive surface PF and cognitive verification PF (hypothesis 3a). It was, also, hypothesized that participants in condition A would provide more accurate PF compared to those in condition B (hypothesis 3b).

To test the impact of peer solution quality on the content of provided PF (type and accuracy), One-way ANCOVA tests were conducted for quality of peer solution (near-correct vs. erroneous) as the independent variable, PF types (cognitive surface, cognitive verification, cognitive elaboration, and self-efficacy) and PF accuracy as the dependent variables, and basic geometric knowledge as a covariate.

The impact of peer solution quality on PF type

The assumption of homogeneity of variances was violated only for cognitive verification PF (p = .008). The assumption of homogeneity of regression slopes was met for all variables (p > .05). After controlling for the covariate basic geometric knowledge, no significant impact was found for peer solution quality on the amount of provided cognitive elaboration PF, $F(1,50) = 0.02$, $p = .901$, $\eta_p^2 = .000$, $B_{10} = 0.11$ (moderate evidence for H_0), or on the amount of provided self-efficacy PF, $F(1,50) = 0.41$, $p = .526$, $\eta_p^2 = .008$, $B_{10} = 0.12$ (moderate evidence for H_0). Participants in condition B did not provide significantly more cognitive elaboration PF (M = .43, SD = .14), or self-efficacy PF (M = .02, SD = .04) compared to those in condition A, (M = .42, SD = .18; M = .03, SD = .06 respectively) (Figure 11). Similarly, no significant impact was found for peer solution quality on the amount of provided cognitive surface PF after controlling for basic geometric knowledge, $F(1,50) = .01$, $p = .943$,

η_p^2 = .000, B_{10} =0.11 (moderate evidence for H_0). Wilcox robust ANCOVA also revealed no significant impact of peer solution quality on the amount of provided cognitive verification PF after controlling for basic geometric knowledge, F_{42} = 0.07, *p* = .947, 95% IC [-.14, .13]; F_{43} = 0.24, *p* = .816, 95% IC [-.12, .10]; F_{44} = 0.12, *p* = .905, 95% IC [-.15, .14]; F_{45} = 0.47, *p* = .646, 95% IC [-.20, .15]. Participants in condition A did not provide significantly more cognitive surface PF (*M* = .04, *SD* = .06) or cognitive verification PF (*M* = .35, *SD* = .16) than participants in condition B (*M* = .04, *SD* = .06; *M* = .32, *SD* = .10 respectively) (Figure 11). The covariate basic geometric knowledge was not related to cognitive elaboration PF, *F*(1,50) = 0.83, *p* = .366, η_p^2 = .016, B_{10} = 0.40 (anecdotal evidence for H_0), self-efficacy PF, *F*(1,50) = 0.30, *p* = .584, η_p^2 = .006, B_{10} = 0.32 (anecdotal evidence for H_0), or cognitive surface PF, *F*(1,50) = 0.77, *p* = .385, η_p^2 = .015, B_{10} = 0.39 (anecdotal evidence for H_0).

The impact of peer solution quality on PF accuracy

According to ANCOVA, a significant impact of peer solution quality was found on PF accuracy after controlling for basic geometric knowledge, *F*(1, 50) = 7.69, *p* = .008, η_p^2 = .133, B_{10} = 124.18 (decisive evidence for H_1) . Participants in condition A (*M* = .53, *SD* = .21) provided significantly more accurate PF than participants in condition B (*M* = .37, *SD* = .20) (Figure 12). The covariate basic geometric knowledge also had a significant impact on PF accuracy, *F*(1, 50) = 10.31, *p* = .002, η_p^2 = .171, B_{10} =19.905 (strong evidence for H_1).

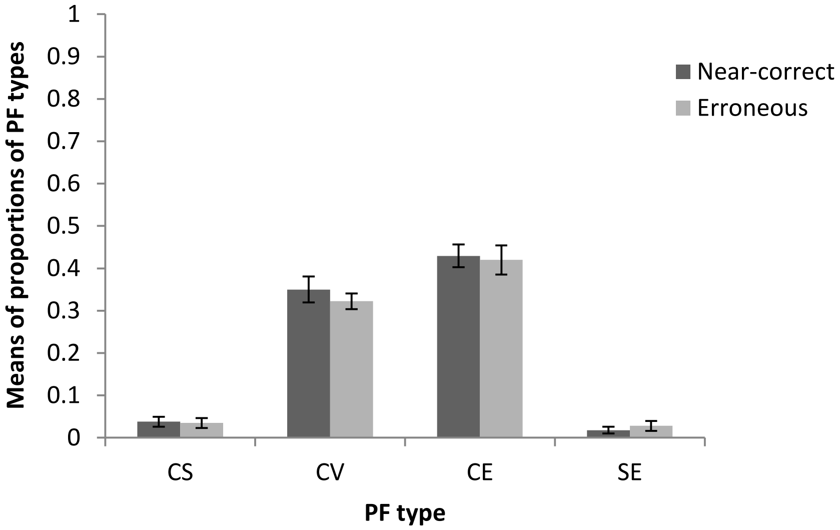

Figure 11. The means of the proportions of PF types cognitive surface (CS), cognitive verification (CV), cognitive elaboration (CE), and self-efficacy (SE) for condition A and condition B. Error bars are standard errors of means.

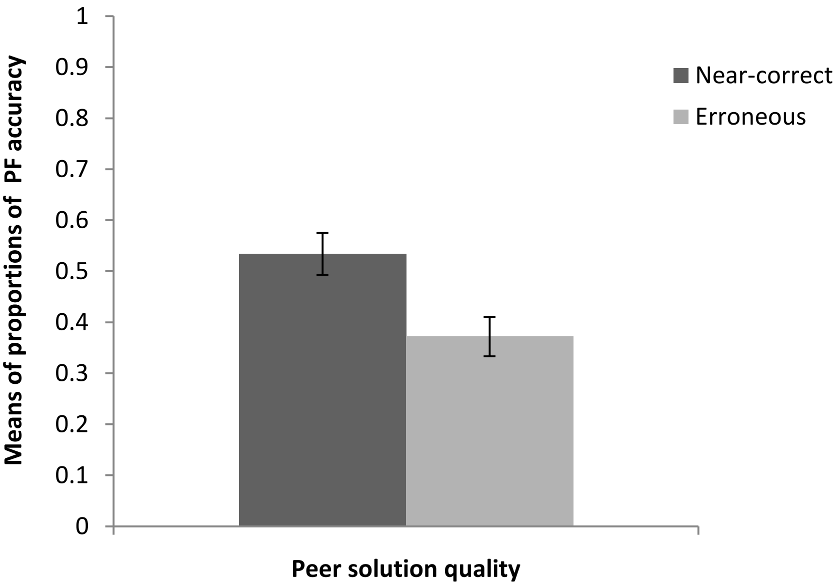

Figure 12. The means of the proportions of PF accuracy for condition A and condition B. Error bars are standard errors of means.

3.3.6. RQ 4: The relationships between the PF providers' PTDT on the peer solution, PF providers' proof comprehension, and content of the provided PF

Regarding the fourth research question, it was hypothesized that the PTDT on the text, figure, and text and figure combined of the peer solution would be associated with PF providers' proof comprehension for both conditions (hypothesis 4a). The PTDT on the text, figure, and text and figure combined of the peer solution was expected to be associated with PF types (cognitive surface, cognitive verification, cognitive elaboration, self-efficacy, and affective) (hypothesis 4b), and with PF accuracy (hypothesis 4c) for both conditions. It was also hypothesized that PF providers' proof comprehension would be positively associated with cognitive elaboration PF and cognitive verification PF; and it would be negatively associated with cognitive surface PF, self-efficacy PF and affective PF (hypothesis 4d). Additionally, PF providers' proof comprehension was also hypothesized to be positively related to PF accuracy for both conditions (hypothesis 4e).

To investigate the relationships between the PTDT on different parts and components of the peer solution, the PF providers' proof comprehension, and the content (type and accuracy) of provided PF, correlations were conducted for each experimental condition separately. Due to violations of the normality assumption by the PF type variables and some of the PTDT measures, Spearman's correlations were used.

No significant correlations were found between PTDT measures (text, figure, and text and figure combined) and proof comprehension for both conditions (p > .05; see Table 5). Similarly, no significant relationships were found between PTDT measures (text, figure, and text and figure combined) and PF types: cognitive surface PF, cognitive elaboration PF, self-efficacy PF, and affective PF, for condition B (p > .05; see Table 6). There was a significant moderate positive correlation between PTDT on the text of the non-comparable peer solution part and cognitive verification PF for condition B, $r_s(24) = .40$, $p = .043$ (see Table 6). For condition A, the PTDT on the text of the non-comparable peer solution part was moderately negatively related to self-efficacy PF, $r_s(25) = -.46$, $p = .015$, whereas, the PTDT on the figure of the non-comparable peer solution part was moderately positively related to self-efficacy PF, $r_s(25) = .46$, $p = .015$ (see Table 6). The PTDT on the text of the comparable peer solution part was moderately negatively related to cognitive elaboration PF, $r_s(25) = -.51$, $p = .007$, and PF accuracy, $r_s(25) = -.43$, $p = .027$ for condition A. Conversely, the PTDT on the figure of the comparable peer solution part was moderately positively related to cognitive elaboration PF, $r_s(25) = .51$, $p = .007$, and to PF accuracy, $r_s(25) = .43$, $p = .027$ (see Table 6).

Table 5

Spearman's correlations between PTDT on different parts and components of peer solution and proof comprehension

	PTDT					
	Non-comparable peer solution part			Comparable peer solution part		
	Text and figure	Text	Figure	Text and figure	Text	Figure
Proof comprehension						
Condition A	-.33	.30	-.30	.33	-.13	.13
p value	.096	.128	.128	.096	.512	.512
Condition B	.06	.32	.00	-.06	.13	-.13
p value	.787	.114	.991	.787	.522	.522

Note. PTDT = proportional total dwell time

Table 6

Spearman's correlations between PTDT on different parts and components of peer solution and PF types and accuracy

	PTDT											
	Non-comparable peer solution part						Comparable peer solution part					
	Text and figure		Text		Figure		Text and figure		Text		Figure	
PF content / Condition	A	B	A	B	A	B	A	B	A	B	A	B
Cognitive surface	.24	-.18	.25	-.09	-.25	-.04	-.24	.18	.21	-.14	-.21	.14
p value	.226	.390	.207	.664	.207	.862	.226	.390	.297	.499	.297	.499
Cognitive verification	.12	-.23	.20	**.40***	-.20	-.22	-.12	.23	.33	.31	-.33	-.31
p value	.543	.266	.312	.043	.312	.274	.543	.266	.088	.127	.088	.127
Cognitive elaboration	-.21	.20	.09	.09	-.09	.23	.21	-.19	**-.51****	-.03	**.51****	.03
p value	.305	.329	.653	.652	.653	.254	.305	.329	.007	.889	.007	.889
Self-efficacy PF	.05	.06	**-.46***	-.05	**.46***	.05	-.05	-.06	-.11	-.17	.11	.15
p value	.792	.761	.015	.822	.015	.804	.792	.761	.590	.476	.590	.476
Affective PF	.01	-	.19	-	-.19	-	-.01	-	-.04	-	.04	-
p value	.954	-	.343	-	.343	-	.954	-	.830	-	.830	-
PF accuracy	-.36	-.17	.12	.06	-.12	-.06	.36	.17	**-.43***	-.02	**.43***	.02
p value	.066	.409	.562	.756	.562	.783	.066	.409	.027	.912	.027	.912

Note. * $p < .05$; PTDT = proportional total dwell time; Condition A = almost-correct peer solution, Condition B = erroneous peer solution, affective PF was not provided in condition B

The correlation analyses between PF providers' proof comprehension and PF types revealed no significant relationships between proof comprehension and cognitive surface PF, cognitive verification PF, cognitive elaboration PF, or affective PF for both conditions ($p >$ .05; see Table 7). However, proof comprehension was weakly negatively correlated with self-efficacy PF for condition A, $r_s(25) = -.39$, $p = .044$, but not for condition B, $r_s(24) = -.02$, $p =$ .923 (see Table 7). PF providers' proof comprehension and PF accuracy were strongly positively correlated for condition A, $r_s(25) = .69$, $p = .000$, and moderately positively correlated for condition B, $r_s(24) = .54$, $p = .004$ (see Table 7).

Table 7

Spearman's correlations between proof comprehension, and PF types and accuracy

	PF types					PF accuracy
	CS	CV	CE	Affective	SE	
Proof comprehension						
Condition A	.21	-.09	.11	.01	**-.39***	**.69*****
p value	.305	.669	.575	.960	.044	.000
Condition B	-.04	.09	.07	-	-.020	**.54****
p value	.838	.660	.722	-	.932	.004

Note. *$p < .05$, ** $p < .01$, *** $p < .001$; Condition A = almost-correct peer solution, Condition B = erroneous peer solution; CS = cognitive surface, CV = cognitive verification, CE = cognitive elaboration, SE = self-efficacy PF.

3.3.7. RQ 5: The impact of the quality of the peer solution (near-correct vs. erroneous) to a geometry proof on the experience of the epistemic emotions curiosity and confusion

The quality of peer solution was expected to influence how PF providers experience curiosity and confusion while providing PF. More specifically, participants in condition B were expected to experience more curiosity and more confusion during PF provision (hypothesis 5).

To investigate the differential effect of the quality of the peer solution on the epistemic emotions curiosity and confusion, two one-way ANCOVA tests were conducted with curiosity or confusion as the dependent variable, quality of the peer solution as the independent variable, and basic geometric knowledge as a covariate. Levene's tests indicated homogeneity of variances for curiosity *(p* = .743) and confusion (*p* = .563). The assumption of homogeneity of regression slopes was met for both emotions (*p* > .05). There was a significant effect of the quality of the peer solution on experienced curiosity, after controlling for basic geometric knowledge, $F(1, 50) = 4.64$, $p = .036$, $\eta_p^2 = .085$, $B_{10} = 0.55$ (anecdotal evidence for H_0). The basic geometric knowledge, $F(1,50) = 0.55$, $p = .463$, $\eta_p^2 = .011$, $B_{10} = 0.31$ (anecdotal evidence for H_0), had no significant effects on curiosity. As illustrated by Figure 13, participants in condition B (*M* = 3.62, *SD* = 0.80) were significantly more curious than those in condition A (*M* = 3.15, *SD* = 0.82). Conversely, there was no significant effect of the quality of the peer solution on confusion, after controlling for basic geometric knowledge, $F(1, 50) = 0.09$, $p = .760$, $\eta_p^2 = .002$, $B_{10} = 0.19$ (moderate evidence for H_0) (Figure 13). The basic geometric knowledge had no significant effect on confusion, $F(1,50) = 2.08$, $p = .156$, $\eta_p^2 = .040$, $B_{10} = 0.64$ (anecdotal evidence for H0). Participants in condition B (*M* = 2.92, *SD* = 0.94) were not significantly more confused than those in condition A (*M* = 2.96, $SD_A = 0.94$) (Figure 13).

Figure 13. The means of experienced curiosity and confusion for condition A and condition B. Error bars are standard errors of means.

3.3.8. RQ 6: The relationships between the PF providers' epistemic emotions, PTDT on the peer solution, proof comprehension, beliefs about PF provision, perceptions of their PF, and the content of their PF

The relationship between PF providers' experienced curiosity, confusion, anxiety, PTDT on the peer solution, proof comprehension, and the content of their PF

It was hypothesized that the PTDT on the text, figure, and text and figure combined of the peer solution would be positively associated with curiosity and confusion for both conditions (hypothesis 6a). PF providers' proof comprehension was expected to be positively associated with curiosity (hypothesis 6b) and to be negatively associated with confusion for both conditions (hypothesis 6c). It was also hypothesized that PF types (cognitive surface, cognitive verification, cognitive elaboration, self-efficacy, and affective) and PF accuracy would be associated with curiosity (hypothesis 6d) and with confusion (hypothesis 6e). Negative correlations were expected to be observed between cognitive elaboration PF, PF accuracy and anxiety, and positive correlations were expected to be observed between self-efficacy PF and anxiety (hypothesis 6f).

The relationships between curiosity, confusion, anxiety, PTDT, proof comprehension, and PF content (type and accuracy) were investigated through Spearman's correlations. No significant relationships were found for curiosity and confusion with PTDT on the text, figure, and text and figure combined of the peer solution for both conditions ($p > .05$; see Table 8).

There were no significant correlations between proof comprehension and curiosity for condition A, $r_s(25) = .33$, $p = .096$, or condition B, $r_s(24) = .08$, $p = .686$ (see Table 9). Proof comprehension was strongly negatively correlated with confusion for condition A, $r_s(25) = -.65$, $p = .000$, but not for condition B, $r_s(24) = .38$, $p = .059$ (see Table 9).

No significant relationships were found between curiosity and PF types (cognitive surface, cognitive verification, cognitive elaboration, self-efficacy, and affective) or PF accuracy for both conditions ($p > .05$; see Table 10). A significant but weak positive relationship was observed between confusion and affective PF for condition A, $r_s(25) = .42$, $p = .029$. Confusion was moderately negatively related to PF accuracy for condition A, $r_s(25) = -.47$, $p = .010$. No significant relationship was found between confusion and the PF types (cognitive surface, cognitive verification, cognitive elaboration, and self-efficacy) for both conditions ($p > .05$; see Table 10). Similarly, no significant correlation was found between confusion and PF accuracy for condition B, $r_s(24) = .19$, $p = .346$.

Table 8

Spearman's correlations between PTDT on different parts and components of peer solution and curiosity and confusion

	PTDT											
	Non-comparable peer solution part						Comparable peer solution part					
	Text and figure		Text		Figure		Text and figure		Text		Figure	
Condition	A	B	A	B	A	B	A	B	A	B	A	B
Curiosity	-.25	.24	-.08	.10	.08	-.15	.25	-.24	-.06	.26	.06	-.26
p value	.206	.242	.694	.632	.694	.468	.206	.242	.786	.192	.786	.192
Confusion	.04	.07	-.29	-.02	.29	-.01	-.04	-.07	-.13	-.14	.13	.14
p value	.830	.721	.144	.910	.144	.977	.830	.721	.525	.485	.525	.485

Note. PTDT = proportional total dwell time; Condition A = almost-correct peer solution, Condition B = erroneous peer' solution

Table 9

Spearman's correlations between proof comprehension, and curiosity and confusion

	Proof comprehension	
	Condition A	Condition B
Curiosity	.33	.08
p value	.096	.686
Confusion	**-.65*****	.38
p value	.000	.059

Note. **p* < .05, ** p < .01, *** p < .001; Condition A = almost-correct peer solution, Condition B = erroneous peer solution

Table 10

Spearman's correlations between curiosity, confusion, anxiety, and PF types and accuracy

	PF types					PF accuracy
	CS	CV	CE	Affective	SE	
Curiosity						
Condition A	.11	-.12	-.13	-.32	.05	.32
p value	.578	.545	.514	.109	.797	.105
Condition B	.19	.09	.08	-	-.26	.14
p value	.346	.651	.690	-	.197	.499
Confusion						
Condition A	-.31	.19	-.12	**.42***	-.20	**-.47***
p value	.122	.357	.537	.029	.323	.010
Condition B	-.03	.06	-.08	-	-.14	.19
p value	.888	.790	.682	-	.512	.346
Anxiety						
Condition A	-	-	-.05	-	.06	**-.49****
p value	-	-	.797	-	.787	.010
Condition B	-	-	-.10	-	.04	-.38
p value	-	-	.632	-	.839	.053

Note. **p* < .05, ** p < .01, *** p < .001; CS = cognitive surface, CV = cognitive verification, CE = cognitive elaboration, SE = self-efficacy PF.

No significant relationships were found between anxiety and cognitive elaboration PF for condition A, $r_s(25) = -.05$, $p = .797$, or condition B $r_s(24) = -.01$, $p = .632$ (see Table 10). Anxiety was moderately negatively related to PF accuracy for condition A, $r_s(25) = -.49$, $p = .010$, but not significantly related to PF accuracy for condition B, $r_s(24) = -.38$, $p = .053$. Similarly, no significant relationship was found between anxiety and self-efficacy PF for condition A, $r_s(25) = .06$, $p = .788$, or condition B, $r_s(24) = .04$, $p = .838$ (see Table 10).

The relationship between PF providers' beliefs about PF provision, their perceptions of their PF, and the content of their PF

Regarding the relationship between beliefs about PF provision and PF content, it was hypothesized that the beliefs about PF provision (LPF, CPF, and RPF) would be positively related to cognitive verification PF, and cognitive elaboration PF, and would be negatively related to self-efficacy PF (hypothesis 6g). Positive correlations were also expected to be observed between beliefs about PF provision (LPF, CPF, and RPF) and PF accuracy for both conditions (hypothesis 6h). Regarding the relationship between PF providers' perceptions of their PF (quality, accuracy, and usefulness) and the accuracy of the provided PF and CPF beliefs, it was hypothesized that PF providers' perceptions of their PF would be positively related to PF accuracy and CPF for both conditions (hypothesis 6i).

No significant correlations were found between beliefs about PF provision (LPF, CPF and RPF) and cognitive verification PF for both conditions ($p > .05$; see Table 11). Similarly, no significant correlations were found between cognitive elaboration or self-efficacy PF, and beliefs about PF provision (i.e., LPF, CPF, and RPF) for both conditions ($p > .05$; see Table 11). Additionally, no significant relationships were found between PF accuracy and LPF and RPF for both conditions (p > .05; see Table 11). However, PF accuracy and CPF were moderately positively correlated for condition A, $r_s(25) = .48$, $p = .012$, and for condition B, $r_s(24) = .45$, $p = .020$ (see Table 11).

Table 11

Spearman's correlations between beliefs about PF provision, and PF types and accuracy

	PF types			PF accuracy
PF beliefs	CV	CE	SE	
LPF				
Condition A	-.34	.14	.23	-.04
p value	.082	.475	.244	.846
Condition B	.12	-.16	.20	-.20
p value	.565	.436	.332	.318
CPF				
Condition A	-.23	.16	-.15	**.48******
p value	.260	.426	.450	.012
Condition B	.06	.18	-.32	**.45***
p value	.756	.392	.108	.020
RPF				
Condition A	-.26	.14	.12	.08
p value	.190	.495	.568	.702
Condition B	.03	.05	.09	.11
p value	.902	.828	.647	.610

Note. *p < .05, ** p < .01, *** p < .001; CV = cognitive verification, CE = cognitive elaboration, SE = self-efficacy PF; LPF = learning from PF provision, CPF = confidence regarding PF provision, RPF = engaging in reasoning during PF provision.

Regarding the relationship between PF accuracy and PF providers' perceptions of their PF (quality, accuracy, and usefulness), moderate positive correlations were found for condition A: quality of own PF, $r_s(25) = .66$, $p = .000$, accuracy of own PF, $r_s(25) = .58$, $p = .001$, and usefulness of own PF, $r_s(25) = .44$, $p = .021$ (see Table 12). However, no significant correlations were found for condition B between PF accuracy and PF providers' perceptions of their PF:

quality of own PF, $r_s(24) = .33$, $p = .104$, accuracy of own PF, $r_s(24) = .25$, $p = .228$, and usefulness of own PF, $r_s(24) = .08$, $p = .710$ (see Table 12).

Furthermore, significant moderate to strong correlations were found between confidence regarding PF provision beliefs (CPF) and PF providers' perceptions of their PF for condition A, quality of own PF, $r_s(25) = .55$, $p = .003$; accuracy of own PF, $r_s(25) = .57$, $p = .002$; usefulness of own PF, $r_s(25) = .53$, $p = .005$, and for condition B, quality of own PF, $r_s(24) = .49$, $p = .012$; accuracy of own PF, $r_s(24) = .70$, $p = .000$; usefulness of own PF, $r_s(24) = .60$, $p = .001$(see Table 12).

Table 12

Spearman's correlations between PF providers' perceptions of their PF, and PF accuracy and confidence regarding PF provision beliefs

Perceptions of PF	PF accuracy	CPF
Quality		
Condition A	**.66*****	**.55****
p value	.000	.003
Condition B	.33	**.49****
p value	.104	.012
Accuracy		
Condition A	**.58****	**.57****
p value	.001	.002
Condition B	.25	**.70*****
p value	.228	.000
Usefulness		
Condition A	**.44***	**.53****
p value	.021	.005
Condition B	.08	**.60****
p value	.710	.001

Note. $*p < .05$, $** p < .01$, $*** p < .001$; CPF = confidence regarding PF provision

3.4. Discussion

The aim of this study was to investigate the differential impact of the quality of peer solutions to a geometry proof on (a) how preservice mathematics teachers who provided PF on these solutions focus their attention on different components of the peer solution (i.e., process the proof), (b) the content of the provided PF, and (c) experienced epistemic emotions curiosity and confusion. Additionally, this study explored the relationship between PTDT on different components of the peer solution (i.e., text, figure, and text and figure combined), PF providers' proof comprehension, content of provided PF, experienced epistemic curiosity, confusion and anxiety, and PF providers' beliefs about PF and their perceptions their PF. Participants were asked to provide verbal PF on one of two variations of a fictional peer solution to the same geometry proof (i.e. a near-correct or an erroneous peer solution). While doing so their eye-movements were recorded using eye-tracking methodology.

3.4.1. RQ 1: The impact of peer solution quality on reading behavior measured by PTDT

There was a significant effect of the quality of the peer solution (near-correct vs. erroneous) on the PTDT on different components of the peer solution. Participants who provided PF on the erroneous peer solution (condition B) spent significantly more time on the text and the figure and on text and figure combined while providing PF on the earlier non-comparable peer solution part compared to those who provided PF on the near-correct solution (hypothesis 1a was supported). Since the figure contains more properties that are not specifically stated in the premises (Koedinger & Anderson, 1991), it seems that the participants needed to check the correctness of statements against the figure in the absence of warrants in condition B.

For the later comparable peer solution part, participants in condition B had significantly more PTDT on the text, whereas participants in condition A had significantly more PTDT on the figure (hypothesis 1b was supported). Encountering more errors earlier in the fictional peer solution – more specifically missing warrants – stimulated an analytical approach to validate the rest of the peer solution. Therefore, participants in condition B spent significantly more time on the text of the comparable peer solution part while providing PF. Conversely, participants who encountered no errors and more warrants earlier in the fictional peer solution (i.e., condition A) seemed to remain with a figure-based approach to validate the remainder of the peer solution. By adopting the figure-based approach participants in condition A did not need to gaze at the text as often as their counterparts, and rather focused on the figure while provide PF on the fictional peer solution. Thus, it might be the case that mathematics preservice teachers rely on figure-based inferences and that the errors or missing warrants may lead them to focus more on the text component. These findings support the assumptions of Ufer et al's (2009) mental models framework of geometry proof competency, and suggest that the two modes of reasoning (analytical and figure-based) can also be observed when students provide PF on peer solutions to geometry proofs. Yet, the quality of the peer solution seems to influence which reasoning approach is adopted. An alternative explanation for the findings might be that the participants in condition B spent longer time on the figure (compared to participants in condition A) when providing PF on the non-comparable peer solution part to test the unwarranted claims. Consequently, these participants did not have to spend a long time on the figure when providing PF on the later comparable peer solution part. This is in line with evidence from text reading research that fixation durations decrease on words observed earlier in the same text when students read them again (Raney & Rayner, 1995). The same pattern might apply to figures in geometry proofs; however, this claim should be tested empirically in future research.

3.4.2. RQ 2: The impact of peer solution quality on PF providers' proof comprehension

The quality of the peer solution had a significant effect on PF providers' proof comprehension. Participants in condition A had a better comprehension of the geometry proof compared to those in condition B, even when their basic geometric knowledge was controlled for (hypothesis 2 was supported). This finding suggests that errors in peer solution may inhibit PF providers' comprehension of the proof. This result is consistent with a previous finding from research on erroneous worked examples that showed that middle-school students did not benefit from studying erroneous worked examples when learning decimals (Isotani et al., 2011). Although several other studies reported that studying erroneous worked examples resulted in cognitive gains for students with high and low domain knowledge (e.g., Groß & Renkl, 2007; Siegler, 2002; Tsovaltzi et al., 2010), the participants in our study did not seem to better comprehend the proof when they provided PF on the erroneous peer solution. One reason might be that the task used in our study is a geometry proof, which is challenging for students to validate (Reiss et al., 2000). Furthermore, unlike our study, studies that reported a positive effect of erroneous worked examples often combined the worked example with an additional instructional support such as self-explanation (Groß & Renkl, 2007; Siegler, 2002) or error detection help through feedback (Tsovaltzi et al., 2010). It might be that preservice mathematics teachers only benefit from providing PF on erroneous geometry proofs over repeated exposure to such activity or with the help of additional instructional support, but this should be tested in future research.

3.4.3. RQ 3: The impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the content of the provided PF

The type of PF that participants provided (cognitive surface, cognitive verification, cognitive elaboration, and self-efficacy) did not differ depending on the quality of peer solution, after controlling for basic geometric knowledge (hypothesis 3a was rejected). This finding is consistent with Patchan and Schunn (2015) who found that the amount of different types of PF (e.g., verification, suggestions for improvement, and criticism) did not differ for high-quality and low-quality academic texts by peers. However, whereas Patchan and Schunn (2015) did not determine the accuracy of PF, the current study revealed that PF provided by preservice mathematics teachers who provided PF on the near-correct peer solution to a geometry proof was more accurate compared to those who provided PF on the erroneous peer solution (hypothesis 3b was supported). This supports previous findings from studies with high school students (grades 7 and 13) (e.g., Reiss et al., 2000; Klieme et al., 2003) and undergraduate mathematics students (e.g., Inglis & Alcock, 2012) which showed that validating erroneous proofs was more challenging for students compared to validating correct proofs. A similar finding was also reported by Zerr and Zerr (2011), as undergraduates in their study appeared to be more successful at assessing the correct peer solution to a proof than the erroneous peer solution. It seems that even preservice mathematics teachers have difficulties validating erroneous geometry proofs compared to correct proofs, which might be due to their limited experience with validating erroneous proofs (Zerr & Zerr, 2011). Whether preservice mathematics teachers can become better at validating erroneous proofs as they accumulate more experiences with validating such proofs is an issue that requires further investigation.

The manipulation check item on which the participants rated on a scale from 0 to100 the correctness of the fictional peer solution they provided PF on, also did not reveal differences between the two fictional peer solutions. Previous findings in proof research also showed that undergraduates do not differentiate between correct and incorrect proofs (e.g., Inglis & Alcock, 2012; Selden & Selden, 2003). Although a study by Jones and Alcock (2014) showed that undergraduates peer assessment of a calculus test could be valid compared to an expert's assessment (r = .77), even when no assessment criteria were used, this study suggests that peer assessment of complex tasks like geometry proofs might not be as valid.

In sum, the quality of the peer solution seems to influence the accuracy of PF, but no evidence was found in this study for influence on the PF types. These findings highlight the importance of measuring different aspects of PF content (i.e., type and accuracy), as they appear to convey different information about preservice mathematics teachers' PF skills. While the differences in PF accuracy indicated that it is easier to identify correct parts in the proof than to identify errors – which is already reported by previous research (e.g., Reiss et al., 2000) – the type of PF that preservice mathematics teachers provide is likely to depend on how they perceive the correctness of the peer solution. For instance, if the PF providers perceive the peer solution as correct they are more likely to provide cognitive verification PF, while if they perceive the peer solution as wrong, they are more likely to elaborate on their comments. Therefore, it might be the case that since the participants' perceptions of the correctness of the fictional peer solutions they provided PF on did not differ, the type of their PF on the two qualitatively different peer solutions did not differ either.

3.4.4. RQ 4: The relationships between the PF providers' PTDT dwell time on the peer solution, PF providers' proof comprehension, and content of the provided PF

No significant relationships were found for both conditions between PTDT on the text, figure or text and figure combined of the fictional peer solution and proof comprehension, although it was hypothesized that longer PTDT would be associated with better proof comprehension (hypothesis 4a was not supported). It could be the case that some participants spent more time attending to different components of the peer solution but never managed to comprehend the proof (Hyönä, 2010) because longer PTDT can also indicate uncertainty or difficulty in comprehending information (Holmqvist et al., 2011). This issue requires further investigation through accompanying eye-tracking measures with cued-retrospective reporting because this approach provides more information about why a participant attends to specific parts of the task for a specific amount of time (Van Gog et al., 2005; Van Gog & Scheiter, 2010). Another explanation for the lack of correlations might be because the participants could read the peer solution after the PF provision phase while answering the proof comprehension test.

For condition A, there was a moderate positive correlation between PTDT on the figure of the non-comparable peer solution part and self-efficacy PF, and a moderate negative correlation between PTDT on the text of the non-comparable peer solution part and self-efficacy PF. There were, also, moderate negative correlations between PTDT on the text of the comparable peer solution part and cognitive elaboration PF and with PF accuracy. Conversely, for condition A, the PTDT on the figure corresponding to the comparable peer solution part was moderately positively related to both cognitive elaboration PF and PF accuracy. That is, the more time participants in condition A spent focusing on the figure corresponding to the non-comparable peer solution part the more self-efficacy PF they provided, but the more time they spent focusing on the figure corresponding to the comparable peer solution part, the more they provided elaborated PF, and the more accurate their PF was. For condition B, more PTDT spent on the text of the non-comparable peer solution part was moderately associated with more cognitive verification PF, but no relationships were found with PF accuracy for this condition (hypotheses 4b and 4c were partially supported). These findings suggest that adopting a figure-based approach may help preservice mathematics teachers provide more accurate and elaborated PF, because this approach may allow identifying reasoning errors – at least the type of errors in the current study – more easily than the analytical text-based approach. Nevertheless, spending time focusing on the figure corresponding to the non-comparable peer solution part was also associated with comments about PF providers' own ability to provide

PF or to perform the proof (i.e., self-efficacy PF). Future research should examine whether adopting a figure-based approach while providing PF on peer solutions to geometry proofs leads to better PF content.

No significant correlations were found between PF providers' proof comprehension and PF types (cognitive elaboration and cognitive verification); except for a weak negative correlation with self-efficacy PF in condition A; hypothesis 4d could not be supported. In contrary, moderate to strong correlations were found between PF accuracy and PF providers' proof comprehension for both conditions; hypothesis 4e was supported. As anticipated, participants who provided more accurate PF also had a better understanding of the proof they provided PF on. This finding aligns with the results from Van Zundert, Sluijsmans et al. (2012) that task-specific knowledge is a pre-requisite for the accuracy of PF. However, almost no evidence was found in the current study that certain types of PF are related to the comprehension of the proof. It was expected that proof comprehension would be associated with more cognitive verification and cognitive elaborated PF, and with less cognitive surface, affective and self-efficacy PF, because students' domain knowledge can influence the type of PF they provide (Patchan & Schunn, 2015), and consequently their task-specific understanding could be related to the content/type of PF provided. The relationship between task-specific understanding (proof comprehension in our study) and different types of PF should be examined further before drawing any conclusions. It is quite evident, though, that the accuracy of students' PF should be taken into account in PF research because the type of PF might not provide the full picture about a students' PF skill. Finally, it is unclear whether PF type, accuracy or both of them have an essential impact on the PF providers' learning gain.

3.4.5. RQ 5: The impact of the quality of the peer solution (near-correct vs. erroneous) to a geometry proof on the experience of the epistemic emotions curiosity and confusion

The current study revealed that participants who provided PF on the erroneous peer solution (condition B) felt more curious (with a medium effect size) than those who provided PF on the near-correct peer solution (condition A), although the Bayes factor indicated an anecdotal evidence for H_0. No significant differences were observed for confusion (hypothesis 5 was partially supported). The quality of the peer solution, therefore, might have a differential effect on experienced curiosity while providing PF on peer solutions of geometry proofs. This finding is in line with emotions research's findings that curiosity is triggered when students encounter conflicting information (Kang et al., 2009). However, what remains unclear is whether curiosity experienced during PF provision leads students to adopt a figure-based or an analytical validation approach or to engage deeply with the task. Furthermore, the disagreement between the ANCOVA test results and the Bayes factor suggest that the finding regarding curiosity should be treated as exploratory.

3.4.6. RQ 6: The relationships between the PF providers' epistemic emotions, PTDT on the peer solution, proof comprehension, beliefs about PF provision, perceptions of their PF, and the content of their PF

The relationship between PF providers' experienced curiosity, confusion, anxiety, PTDT on the peer solution, proof comprehension, and the content of their PF

No significant relationships were found between curiosity or confusion and PTDT on any component of the peer solution for both conditions (hypothesis 6a was not supported). Curiosity was not significantly related to PF providers' proof comprehension for both conditions (hypothesis 6b was not supported). Confusion was negatively related to PF providers' proof comprehension, but only for condition A (hypothesis 6c was supported). Curiosity and confusion were expected to be associated with PTDT spent on the peer solution and with proof comprehension, as these emotions have been reported to stimulate more task engagement (D'Mello & Graesser, 2012; Muis et al., 2015). Although there was some indication that confusion is related to proof comprehension, this was only the case when the peer solution was near-correct and the correlation was negative. It might be the case that confusion functions differently depending on the quality of the peer solution being validated, and the level of domain knowledge of the PF provider. For example, students with low domain knowledge may experience confusion but may still not be able to resolve the cognitive conflict they experience while providing PF on complex tasks like geometry proofs.

No significant correlations were found between curiosity and PF types or PF accuracy (hypothesis 6d was not supported). For condition A, affective PF was (weakly) positively correlated with confusion, whereas PF accuracy was (moderately) negatively correlated with confusion. These findings are consistent with the negative correlation between PF providers' proof comprehension and confusion, and show that this emotion appears to be maladaptive at least for condition A. These findings are also consistent with the findings by Muis et al. (2015) that confusion negatively predicted shallow and deep processing during mathematics problem solving. No significant relationships were found in both conditions between confusion and cognitive surface PF, cognitive verification PF, cognitive elaboration PF, or self-efficacy PF, thus hypothesis 6e was only partially supported. A significant negative relationship was found between anxiety and PF accuracy (only for condition A), but no significant relationship was found between anxiety and cognitive elaboration PF or self-efficacy PF for both conditions (hypothesis 6f was partially supported). This study shows that the accuracy of PF might be related to confusion, as well as anxiety – an emotion reported as experienced by students in PF research (e.g., Cartney, 2010; Cheng, Hou, & Wu, 2014; Harris & Brown, 2013). These findings indicate a need to further investigate the relationships between confusion and anxiety and PF types.

The relationship between PF providers' beliefs about PF provision, their perceptions of their PF, and the content of their PF

Beliefs about PF provision (LPF, CPF, and RPF) were not related to PF types (cognitive verification, cognitive elaboration or self-efficacy) for both conditions (hypothesis 6g was not supported). Nevertheless, beliefs related to confidence regarding PF (CPF) were positively related to PF accuracy in both conditions. Students who were accurate in their PF also held more positive CPF beliefs. This finding highlights the importance of this type of beliefs which according to the theory of planned behavior (Ajzen, 2005) can influence individual's motivation to perform a behavior. No significant relationships were found between PF accuracy and LPF, or RPF (hypothesis 6h was partially supported). Furthermore, PF providers' perceptions of their PF (quality, accuracy, and usefulness) were moderately to strongly positively related to PF accuracy but only for condition A, and were moderately to strongly related to CPF for both conditions (hypothesis 6i was partially supported). These findings might indicate that preservice mathematics teachers who provide accurate PF are also more likely to be aware that their PF is of good quality, accurate, and useful for their peer. Yet, the quality of peer solution may play a role in these relationships.

3.4.7. Methodological limitations

Although no significant relationships were found between the eye-tracking measures and PF providers' comprehension of the proof, we gained more insight into how preservice mathematics teachers distribute their locus of attention on different components of the proof (text and figure) depending on the quality of the peer solution they are providing PF on. However, using multiple eye-tracking measures to examine how students process the learning task is essential (Rayner, 1998). Especially with geometry proofs that usually contain a figural component, measuring the number of transitions from the text to the figure component of the proof and vice versa can provide information about how both components of the tasks are integrated. Moreover, cued-retrospective reports can provide answers to questions about failure(s) in processing (Hyönä, 2010), and whether longer PTDT is a valid measure of cognitive processing while providing PF.

The absence of a relationship between PTDT on different components of the peer solution and PF providers' proof comprehension might be due to the speed of the eye-tracker used in the current study (25 Hz). The low speed might have produced more pronounced measurement error that could be reduced by using larger sample sizes. Nevertheless, the current study signifies a first attempt to examine cognitive processing while providing PF. Thus the findings should be replicated with a larger sample size, with multiple types of eye-measures, and using higher recording speed to test whether a relationship between PTDT and proof comprehension exists. Individual testing usually required in eye-tracking research, and the amount of data produced, create consistent challenges for researchers to conduct studies with large sample sizes. One solution to increase the power of future studies is to use within-subject design that will also reduce the effects of individual variability in eye-tracking measures (Holmqvist et al., 2011). Within-subject designs are, however, not always suitable when the target is to investigate the cognitive processing of a task because the frequency of exposure to the task changes how students fixate on different components of the task (see Raney & Rayner, 1995).

This study examined the epistemic emotions curiosity, confusion and anxiety, assuming that these emotions are activated while providing PF on peer solutions to complex learning tasks like geometry proofs, and might influence the comprehension of the proof and the PF content. Nevertheless, social emotions (e.g., shame, envy, and sympathy) are, also, expected to take part in PF activities, although their role is yet to be explored. Finally, this study explored PF providers' cognitive processing while providing PF on a peer solution to a geometry proof, which is a challenging task for university students including preservice mathematics teachers; hence the findings might not be generalizable to other simpler tasks. Employing an eye-tracking methodology to investigate cognitive processing during PF provision in other domains, using tasks of different difficulties remains an open area for future investigation.

3.4.8. Practical implications

The findings of the current study suggest that as a result of providing PF on peer solutions to geometry proofs, preservice mathematics teachers are more capable of comprehending the proof and providing more accurate PF when they provide PF on a near-correct peer solution. Therefore, when employing PF activities to teach geometry proofs to preservice teachers or to train their proof validation skills, instructors should start with easy tasks (to ensure having more correct peer solutions), or start by providing solutions with few well-focused errors and gradually introduce more erroneous solutions as students become more competent assessors. This way, preservice mathematics teachers' proof validation skills – as a learning goal – can be developed without harming their understanding of the proofs.

Furthermore, as found in the current study and prior research, there is evidence that even undergraduates do not differentiate between correct and incorrect proofs (Inglis & Alcock, 2012; Selden & Selden, 2003). Consequently, instructors are advised to refrain from assigning an official weight (e.g., as part of grading) to scores and comments by peers on geometry proofs in teacher-training classes. Finally, anxiety and confusion appear to be associated with less accurate PF, suggesting that instructors should reduce the levels of these emotions experienced during PF provision. Preparing preservice teachers for the PF activity through open discussions about the purpose of PF as a learning tool (Molloy et al., 2013), welcoming errors in the classroom (Borasi, 1994), and training preservice teachers to provide PF (Sluijsmans et al., 2004), may all help to reduce anxiety associated with PF activities. Furthermore, providing more instructional support (e.g., worked examples) during PF activities might help to reduce confusion experienced by preservice mathematics teachers while providing PF.

4. General discussion

Peer feedback (PF) is considered as one of the key practices of assessment for learning (AfL), because students take an active role in assessing their learning, and feedback is a major product of this activity (Black & Wiliam, 2009). In mathematics education, proof validation is regarded as inseparable skill from proof comprehension and proof construction (Selden & Selden, 2015a). Research investigating the role of PF in improving students' assessment skills of proofs such as proof validation is still limited (e.g., Lavy & Shiriki, 2014; Zerr & Zerr, 2011). PF is of particular importance for preservice mathematics teachers since they need assessment skills not only to assess their performance in the future but also to assess the performance of their students and peers.

In some related domains to PF, such as peer tutoring, evidence suggests that tutors learn from tutoring more than tutees (for a review see Graesser, D'Mello, & Cade, 2011). However, learning gains from PF provision are still hardly studied. There is some evidence that students' performance can improve as a result of providing PF (Cho & MacArthur, 2011). Several theoretical assumptions were proposed as explanations of the cognitive processes underlying learning gains from PF provision (Cho & MacArthur, 2011; Nicol et al., 2014; Topping & Ehly, 2001). However, to date, very few studies investigated the cognitive processes involved in PF activities (e.g., Bolzer et al., 2015).

As was illustrated in the M^{2}IPA framework introduced in Chapter 1, a central process to PF provision is the PF composition process. Whereas most of the PF research focused on the PF message (e.g., Gan & Hattie, 2014; Gielen et al., 2010; Guasch et al., 2013; Walker, 2015), the PF composition process received little to no attention. Moreover, the PF composition process is influenced by providers' individual characteristics (e.g., domain knowledge, beliefs, and emotions) and the representation of the recipient (e.g., level of knowledge, or sensitivity to feedback) (Strijbos & Müller, 2014). These individual characteristics, and how they shape the PF composition process and the PF message, also received little attention in research about PF.

This dissertation intends to investigate the mechanisms and benefits of PF provision by investigating PF provision on peer solutions to mathematics tasks that require scientific reasoning and argumentation while considering: the content of the PF message, the PF composition process, and the relationships between the two components. The role of students' individual characteristics was also explored, including domain knowledge, epistemic emotions, beliefs about PF provision, and PF providers' perceptions of their PF. Two empirical studies were conducted with preservice mathematics teachers. The first study focused on training PF provision skills, whereas the second study employed an eye-tracking methodology to examine the PF composition process. The following research questions were investigated for each empirical study:

Study 1:

RQ 1. What is the impact of a structured PF provision training on preservice mathematics teachers' PF provision skills, and will students with different levels of domain knowledge benefit differentially from the training?

RQ 2. Will preservice mathematics teachers' beliefs about PF provision change after the training, and will domain knowledge play a role in that change?

Study 2:

RQ 1. How does cognitive processing while providing PF (measured by proportional total dwell time) differ depending on the quality of a peer solution (near-correct vs. erroneous) to a geometry proof?

RQ 2. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the PF providers' comprehension of the proof?

RQ 3. What is the impact of the peer solution quality (near-correct vs. erroneous) to a geometry proof on the content of the provided PF?

RQ 4. What are the relationships between the PF providers' proportional total dwell time on the peer solution, PF providers' proof comprehension, and content of the provided PF?

RQ 5. What is the impact of the quality of the peer solution (near-correct vs. erroneous) to a geometry proof on the experience of the epistemic emotions curiosity and confusion?

RQ 6. What are the relationships between the PF providers' epistemic emotions, proportional total dwell time on the peer solution, proof comprehension, beliefs about PF provision, perceptions of their PF, and the content of their PF?

In the following sections, the main findings from the empirical studies (reported in Chapters 2 and 3) will be summarized, followed by a synthesis of both studies. Then, the limitations of the two studies will be addressed, before discussing their theoretical and practical implications. Finally, directions for future research and concluding remarks will be presented.

4.1. Summaries of studies

4.1.1. Study 1: Training peer feedback skills on geometric construction tasks: Role of domain knowledge and peer feedback levels

The first study was conducted to investigate whether preservice mathematics teachers' PF provision skills on geometric construction tasks could be improved by involving them in a structured PF training. In a field study, a quasiexperimental mixed design was adopted. Forty-three preservice mathematics teachers took part in a course-based PF training activity that ran throughout a 13-week semester. Participants' basic written PF skills, their basic geometric knowledge, and their beliefs about PF provision, were measured at the start of the semester. Based on the participants' performance in the basic geometric knowledge test, they were grouped into three domain knowledge groups: lowest, middle, and highest one-third of the sample.

All participants received PF provision training over two forty five-minute sessions, during which progressive PF provision prompts, based on Hattie and Timperley's (2007) feedback model, and an evaluation rubric were introduced. The preservice mathematics teachers then practiced providing written PF over two practice sessions with the consultation of the PF provision prompts, the evaluation rubric, and a worked example. Participants' PF provision skills and their beliefs about PF provision were measured again at the end of the semester. In this study, preservice mathematics teachers' written PF was analyzed based on the multi-level model (Hattie & Timperley, 2007) that differentiates between self-level, task-level, process-level, and self-regulation level feedback. Beliefs about PF provision were measured using the PFPQ that consisted of three subscales: learning from PF provision (LPF), confidence regarding PF provision (CPF), and engaging in reasoning during PF provision (RPF).

The findings from Study 1 showed that providing PF at the highest level of Hattie and Timperley's model (i.e., self-regulation) improved after the structured training, but only for preservice mathematics teachers with medium and high domain knowledge. Those with low domain knowledge ended up providing more PF at the task level than the medium domain knowledge preservice mathematics teachers after the training. All three PF provision beliefs decreased significantly after the PF training (with small to medium effect sizes) regardless of preservice mathematics teachers' domain knowledge. It was argued that a basic level of domain knowledge is required for training PF skills, even in the presence of domain knowledge scaffolds (i.e., worked example) to ensure successful outcomes for all preservice mathematics teachers. Further, it was concluded that preservice mathematics teachers' beliefs about PF provision become more realistic with more exposure to PF provision.

4.1.2. Study 2: The impact of peer solution quality on peer feedback provision on geometry proofs: links to peer feedback content, proof comprehension, and individual characteristics

The purpose of the second study was to closely explore the PF composition process, by monitoring preservice mathematics teachers' eye-movements while providing verbal PF on a fictional peer solution to a geometry proof, and by investigating preservice mathematics teachers' individual characteristics likely to influence PF composition. More specifically, we examined the impact of peer solution quality (near-correct vs. erroneous) on: (a) proportional total dwell time (PTDT) on different components of the peer solution to a geometry proof, (b) PF providers' proof comprehension, (c) PF content, and (d) experienced curiosity and confusion. The relationships between PTDT, proof comprehension, PF content, epistemic emotions curiosity, confusion, and anxiety, as well as, beliefs about PF provision and PF providers' perceptions of their PF were also explored. A between-subject experimental design with two conditions was adopted. Fifty-three preservice mathematics teachers were randomly assigned to one of two conditions, in which they had to provide verbal PF either on a near-correct or an erroneous fictional peer solution to a geometry proof. Participants' eye-movements were recorded using a head-mounted eye-tracker during the PF provision phase (on-screen). In the second phase (paper-and-pen), the preservice mathematics teachers answered a proof comprehension test, a basic geometric knowledge test, and they completed epistemic emotions, and PF providers' perceptions of their PF questionnaires. Similar to the first study, the PFPQ was also used to measure students' beliefs about PF provision.

A comparison between both experimental conditions revealed that the eye-movements (PTDT) of those who provided PF on the near-correct peer solution were different from those who provided PF on the erroneous peer solution. While providing PF on the earlier non-comparable peer solution part, students who provided PF on the erroneous peer solution spent more PTDT on both the text and the figure of the peer solution. However, in the later comparable peer solution part, the students who provided PF on the near-correct peer solution spent more PTDT on the figure, whereas those who provided PF on the erroneous peer solution spent more PTDT on the text of that peer solution part. Furthermore, although PTDT did not correlate with the proof comprehension scores, students who provided PF on the near-correct peer solution outperformed those who provided PF on the erroneous peer solution at the proof comprehension test. It was concluded that providing PF on a near-correct peer solution might facilitate adopting a figure-based approach, whereas providing PF on the erroneous solution stimulates an analytical approach.

No differences were found in the type(s) of provided PF (e.g., cognitive verification, cognitive elaboration, self-efficacy) between both condition. However, the PF provided on the near-correct peer solution was more accurate than that provided on the erroneous peer solution. This difference was attributed to students' better ability to validate correct geometry proofs, compared to erroneous proofs (Healy & Hoyles,1998; Reiss et al., 2000; Sommerhoff et al., 2016). PTDT on the text of the non-comparable peer solution part was negatively moderately related to self-efficacy PF, whereas the PTDT on the figure of the same peer solution part was positively moderately related to self-efficacy PF, but only for the near-correct peer solution condition. Similarly, for the same condition, PTDT on the text of the comparable peer solution part was negatively moderately related to cognitive elaboration PF and PF accuracy, while the PTDT on the figure of the same peer solution part was moderately positively related cognitive elaboration PF and PF accuracy. For the erroneous peer solution condition, the PTDT on the text of the non-comparable peer solution part was positively moderately related to cognitive verification PF. It was suggested that adopting a figure-based approach may facilitate providing more cognitive elaboration PF more accurate PF, but only when the peer solution to geometry proof is near-correct. Proof comprehension was not significantly related to PF types, yet it was positively related to PF accuracy for both conditions. It was argued that this finding supports the finding by Van Zundert, Sluijsmans et al. (2012) that students require having knowledge about the task to be able to provide accurate PF.

Participants who provided PF on the erroneous peer solution experienced more curiosity than those who provided PF on the near-correct peer solution, but no significant difference was found in confusion. Furthermore, PF accuracy was found to be negatively associated with confusion and anxiety and positively related to confidence regarding PF belief. PF providers' perceptions of their PF were also moderately to strongly related to PF accuracy, but only for those who provided PF on the near-correct peer solution. Preservice mathematics teachers' perceptions of their PF and their confidence regarding PF provision beliefs were moderately to strongly positively related. These findings suggest that preservice mathematics teachers' confidence regarding PF provision and perceptions of their PF, as well as experienced confusion and anxiety, may be indicators of the accuracy of PF they provide.

4.2. A synthesis of the studies

The main aim of this dissertation was to investigate the PF provision process on complex geometric tasks that involve scientific reasoning and argumentation. More specifically, the two empirical studies examined one or more preservice mathematic teachers' individual characteristics, including: (a) domain knowledge of the PF provider or recipient, (b) epistemic emotions, (c) beliefs about PF provision, and (d) PF providers' perceptions of their PF. The relationship between individual characteristics and the content of the PF message was also investigated. Although both studies focused mainly on PF provision, they differed in their designs, focus, and how the content of the PF message was operationalized (Table 13 summarizes similarities and contrasts between both studies). In Study 1 (Chapter 2), a PF training was implemented to improve preservice mathematics teachers' PF provision skills using domain knowledge and PF provision scaffolds. The content of the PF message was coded based on Hattie and Timperley's (2007) model, which was also used as a basis for the training. Since previous studies found that, without training, students provide PF mostly at the task-level (e.g., Gan & Hattie, 2014; Harris et al., 2015) Hattie and Timperley's model was not used to code PF in the second study because participants did not receive any PF provision training. Furthermore, when providing PF, students often mention comments regarding their own ability to provide PF (i.e., self-efficacy PF; Strijbos et al., 2012). This type of PF is not accounted for by the Hattie and Timperley model. In Study 2 (Chapter 3), the content of the PF message was instead operationalized in terms of its type (cognitive surface, cognitive elaboration, cognitive verification, self-efficacy and affective), and its accuracy. The geometric tasks used in the two studies were not exactly the same. In Study 1, fictional peer solutions of geometric construction tasks were used, whereas in Study 2 fictional peer solutions of a geometry proof were used. However, both tasks require deductive reasoning to solve them. Due to the differences between the two studies in this dissertation, integration of their findings is only feasible for the overarching aspects. In the following subsections, some general findings from both studies will be discussed regarding the main aims of this dissertation, and some common conclusions will be drawn by integrating the findings from both studies.

Table 13

Comparisons between Study 1 and Study 2

	Study 1	**Study 2**
Main focus	Training PF provision at higher levels (process and self-regulation; Hattie & Timperley, 2007) on peer solutions to geometric construction tasks.	Exploring the PF composition process during PF provision on different qualities of peer solution to geometry proofs (near-correct vs. erroneous).
Design	Quasi-experimental mixed-design: Between-subject factor: Domain knowledge Within-subject factor: Measurement occasion	Between-subject experimental design: Near-correct peer solution condition versus Erroneous peer solution condition
Investigated individual characteristics	(1) PF providers' domain knowledge. (2) PF providers' beliefs about PF provision.	(1) PF recipients' domain knowledge represented by the quality of peer solution (near-correct vs. erroneous). (2) PF providers' experienced epistemic emotions curiosity, confusion, and anxiety. (3) PF providers' beliefs about PF provision. (4) PF providers' perceptions of their PF message (quality, accuracy, and usefulness).
Peer feedback content	PF levels (Hattie & Timperley, 2007) – PF type only.	PF type (Strijbos et al., 2012) and accuracy.

4.2.1. Role of domain knowledge

Domain knowledge is one of the essential students' individual characteristics that can influence PF provision in different ways. Domain knowledge is a pre-requisite for PF provision

(Van Zundert, Sluijsmans et al., 2012), and it can determine the type of PF students provide (Patchan et al., 2013). The domain knowledge of the recipient influences the quality of the peer solution, and is therefore expected to influence the type of PF provided on that solution (Patchan & Schunn, 2015). Study 1 focused on the level of the PF providers' domain knowledge, and its impact on how preservice mathematics teachers benefited from PF training. In Study 2 the focus was instead on the quality of the peer solution (representing the domain knowledge of the recipient), and how it influences the PF composition process and the PF message.

As reported in Study 1, only preservice mathematics teachers with medium and high levels of domain knowledge benefited from the PF training and provided more self-regulation PF after the training. Preservice mathematics teachers with low domain knowledge ended up providing more PF at the basic level (i.e., task-level) compared to their medium domain knowledge counterparts. These findings are consistent with previous research which showed that domain knowledge is essential for training PF provision (Van Zundert, Sluijsmans et al., 2012; Van Zundert, Könings et al., 2012). Even with the help of domain knowledge scaffolds (i.e., worked example) and PF provision scaffolds, preservice mathematics teachers with low domain knowledge could not provide PF at the higher level (i.e., self-regulation). Since self-regulation PF requires looking beyond the task and coming up with self-reflective questions that encourage the recipient to monitor and evaluate their learning (Hattie & Timperley, 2007), having a good knowledge of the task is a basic requirement for this type of PF. It was, therefore, recommended that PF training should be introduced after students develop an adequate level of domain knowledge.

Another way domain knowledge can influence PF provision, is through the quality of the peer solution. In Study 2, it was illustrated that the quality of peer solution (near-correct vs. erroneous) influenced how preservice mathematics teachers cognitively process peer solution to a geometry proof while providing PF on that peer solution. Eye-tracking measures (i.e., PTDT) showed that preservice mathematics teachers who provided PF on the erroneous fictional peer solution appeared to adopt an analytical approach by focusing more on the peer solution's text; whereas those who provided PF on the near-correct peer solution adopted a figure-based approach. Although no significant relationships were found between PTDT on different parts of the peer solution and PF providers' comprehension of the proof, those in the near-correct peer solution condition achieved significantly better comprehension of the proof than students in the erroneous peer solution condition, even after controlling for their basic geometric knowledge. This study adds to the findings of proof validation research (e.g., Selden & Selden, 2003; Inglis & Alcock, 2012) by suggesting that the quality of the validated geometry proof can influence what students focus on during proof validation, as well as their comprehension of the validated proof.

The finding that preservice mathematics teachers in the near-correct peer solution condition comprehended the proof they provided PF on better than those in the erroneous peer solution condition, is consistent with the findings by Isotani et al. (2011) who reported that middle-school students performance on decimals tasks after studying erroneous examples was not better than studying worked examples or being involved in partially-supported problem solving. However, studying erroneous examples was found by another study to be more beneficial than correct examples for students' knowledge about identifying errors or incorrect strategies to the problem (i.e., negative knowledge) (Heemsoth & Heinze, 2014). Although providing PF on a peer solution to proof is assumed to contribute to the providers' understanding of the proof (Zerr & Zerr, 2011), this depends on the quality of the peer solution. It seems that the proof validation action taking place during PF provision per se does not guarantee deeper processing of the proof, as students might still focus on surface features of the proof. Study 2 also revealed that proof comprehension was positively (moderate to strong) related to PF accuracy. This finding illustrates that preservice mathematics teachers require a good understanding of the proof to be able to provide more accurate PF.

As reported in Study 2, the quality of the fictional peer' s solution did not influence the type of provided PF, although it was hypothesized that the erroneous peer solution would elicit more cognitive elaboration PF and self-efficacy PF. This finding is consistent with Patchan and Schunn (2015) who found no effect of the quality of peer's written academic text on the type of provided PF. Nevertheless, Study 2 revealed that the PF provided on the near-correct fictional peer solution to a geometry proof was more accurate than that provided on the erroneous fictional peer solution, after controlling for preservice mathematics teachers' basic geometric knowledge. This finding is in line with previous studies conducted with high-school students, which showed that validating erroneous geometry proofs was more challenging for students than validating correct proofs (Reiss et al., 2000; Klieme et al., 2003). It seems that, even for preservice mathematics teachers, detecting errors in geometry proofs is more challenging than identifying correct parts. However, this finding is inconsistent with the finding by Cho and Cho (2011), that students detect more errors in low-quality academic texts. This inconsistency might be due to the differences in the nature of the task – i.e., written text in their study compared to geometry proof in Study 2.

It can be concluded from both studies in this dissertation that domain knowledge is one of preservice mathematics teachers' individual characteristics that shape their PF experience. Preservice mathematics teachers need a basic level of domain knowledge before being able to provide PF at high levels (i.e., self-regulation), yet the domain knowledge of the PF recipient also appears to influence the experience of the PF provider. The domain knowledge of the PF recipient determines the quality of the peer solution, which in turn can determine how the PF provider approaches the solution, their proof comprehension, and the accuracy of their PF. That is, the domain knowledge of the PF provider is likely to influence the PF experience in a different way than the domain knowledge of the recipient (Cho & Cho, 2011). In fact, the domain knowledge of the PF provider and that of the recipient can be expected to have an interaction effect on how the PF provider goes about the composing of the PF, how the PF message will appear, and what they can possibly learn from PF provision. Based on the findings of the current studies, and findings of studies on learning from erroneous examples (e.g., GRoß & Renkl, 2007; Heemsoth & Heinze, 2014; Isotani et al., 2011) it might be the case that preservice mathematics teachers with high domain knowledge are likely to benefit from providing PF to peers with low domain knowledge (erroneous peer solution). Conversely, low domain knowledge preservice mathematics teachers might benefit more from providing PF to a high domain knowledge peer (i.e., correct peer solution).

4.2.2. Role of beliefs about peer feedback provision, providers' perceptions of their peer feedback, and epistemic emotions experienced by the peer feedback provider

Preservice mathematics teachers' beliefs about PF provision were measured in Studies 1 and 2, because preservice teachers' beliefs are suggested to be indicators of how they learn and how they approach learning tasks in their future classrooms (Richardson, 1996). The beliefs were measured with the PFPQ that covers belief about learning from PF provision, confidence regarding PF provision, and engaging in reasoning during PF provision. Study 1 was concerned with changes in preservice mathematics teachers' beliefs after a PF training. As reported in Study 1, preservice mathematics teachers' beliefs became less positive after the PF training, regardless of the students' level of domain knowledge. This finding is consistent with Wang's (2014) study in which students' perceived usefulness of PF decreased over several engagements in PF activities. The decrease in preservice mathematics teachers' beliefs about PF provision was attributed to students' overestimation of their ability to provide PF, which might have become more realistic after being exposed to several PF provision activities.

Furthermore, Cheng and Warren (1997) recommended to involve students in different activities of PF including defining assessment criteria, and to train them to provide PF to avoid negative belief development. This might be one of the explanations for the inconsistency between the findings of Study 1 and the study by Sluijsmans et al. (2004), in which preservice teachers' perspective about PF changed positively after their intervention. In their study, preservice teachers did not only receive PF training but they were also involved in defining assessment criteria.

In Study 2, we explored the relationship between preservice mathematics teachers' beliefs about PF provision and the content of PF they provided, as well as their perceptions of their PF. Study 2 findings revealed moderate positive correlations between preservice mathematics teachers' confidence regarding PF provision (CPF) and PF accuracy indicating that those who provide accurate PF might have more confidence regarding PF provision. This finding is consistent with previous workplace feedback study that showed that feedback self-efficacy was positively related to performance appraisal (Linderbaum & Levy, 2010). Nevertheless, the CPF beliefs cover not only PF self-efficacy but also beliefs regarding having sufficient knowledge to provide PF. These beliefs were also positively (moderate to strong) related to preservice mathematics teachers' perceptions of the quality, accuracy, and usefulness of their PF. The accuracy of PF was moderately positively related to PF providers' perceptions of their PF – but only for condition A. It might be the case that preservice mathematics teachers who provide more accurate PF also perceive the PF they provide as being of high quality, as more accurate, and as useful for their peers. Yet, this needs further investigation especially that a significant relationship was only found for the near-correct peer solution condition. This finding is consistent with findings from the self-assessment literature that students who are able to perform a task are accurate in judging their own performance on that task (for a review see Panadero et al., in press).

Despite the central role of emotions in assessment (Molloy et al., 2013), and the frequently reported emotional responses by students regarding assessment (Carless, 2006; Harris & Brown, 2013), emotions in PF research are still hardly examined. Study 2 explored epistemic emotions likely to be experienced by preservice mathematics teachers when they encounter conflicting information while they provide PF on geometry proofs performed by (fictional) peers, namely curiosity and confusion and anxiety. As reported in Study 2, preservice mathematics teachers who provided PF on the erroneous fictional peer solution were more curious than those who provided PF on the near-correct solution. This finding supports the epistemic emotions literature which posits that epistemic emotions arise when students encounter conflicting information (e.g., D'Mello & Graesser, 2012; Pekrun & Stephens, 2012; Pekrun et al., 2016). However, epistemic curiosity should stimulate students to engage in higher-order cognitive processing (Muis et al., 2015), and this was not reflected in the proof comprehension or the accuracy of PF of preservice mathematics teachers who provided PF on the erroneous fictional peer solution. This signals – at least in the current study – that curiosity was not associated with better learning outcomes (i.e., understanding of the proof PF accuracy). Additionally, confusion and anxiety were weakly to moderately negatively related to PF accuracy (for the near-correct peer solution condition), which supports the notion that negative emotions are associated with rigid, surface learning strategies (Pekrun, 2011). Further studies should be conducted to further investigate the relationships between these epistemic emotions, proof comprehension, and PF accuracy.

It can be concluded that preservice mathematics teachers' beliefs about PF provision, their perceptions of their PF and epistemic emotions they experience during PF provision might be related to the accuracy of their PF. It is important to stress, though, that these findings are only correlational. They might show some indications that these individual characteristics can influence PF composition, but additional research should be conducted to better understand their roles during PF provision (composition).

4.2.3. Peer feedback content

The content of PF is a major element in research about PF, yet researchers differ in how they operationalize PF content. In Study 1, PF content was operationalized in terms of PF levels adopting Hattie and Timplerley's (2007) feedback model. The reason for this choice was to help preservice mathematics teachers approach the geometric construction task at a deeper level. Instead of examining surface features of a peer solution (e.g., structure of proof, grammatical mistakes), this progressive feedback model can help students to focus on advanced processes to solve the task (i.e., process-level feedback), and monitoring and self-evaluation (i.e., self-regulation level feedback). In Study 2, the content of PF was operationalized along two-dimensions, namely type and accuracy. Since most PF is provided at the task-level (Gan & Hattie, 2014; Harris et al., 2015) when students do not receive PF prompts and/or training, the Hattie and Timperley model was not used in Study 2. Yet, there is an overlap between the feedback levels in Hattie and Timperley's model, and some of the feedback purposes (Strijbos et al., 2012) used to code PF types in Study 2. For instance, task-level PF resembles the cognitive purpose, process and self-regulation levels are similar to the metacognitive purpose, and self-level PF is similar to the affective PF purpose in the Strijbos et al.'s (2012) coding scheme. Consequently, it is still possible to carefully draw general conclusions regarding PF types from Study 1 and Study 2. The PF accuracy was added in Study 2, because, unlike experts, peers can have inaccuracies in the PF they provide and PF type does not capture these inaccuracies.

As reported in Study 1, process-level PF did not significantly improve after the PF training, regardless of preservice mathematics teachers' level of domain knowledge. This finding was inconsistent with Gan and Hattie (2014) who found an increase in PF at the process-level. In Study 2, no metacognitive PF was observed, although this type of PF was identified by Strijbos et al. (2012). One explanation for these findings might be the nature of the tasks used in Studies 1 and 2. In the study by Gan and Hattie (2014), the task was a chemistry lab report, for which the procedures of carrying out an investigation are quite visible, and students can quite naturally provide PF on them. Similarly, the task used by Strijbos et al. (2012) was a bachelor's thesis proposal for which metacognitive comments about general approaches used to conduct the project, and to self-monitor and evaluate own progress can be regarded more common than for geometric tasks. For the tasks used in Study 1 and Study 2, most of the processes and approaches that students adopt are not visible, because for these tasks (and other mathematics tasks in general) students only write the solution and not how they came up with that solution. Most of the processes of performing the proofs or the construction tasks remain implicit, which makes it difficult to provide PF at the process-level or PF of a metacognitive purpose. This can be particularly the case when preservice mathematics teachers receive no PF prompts as a scaffold or no PF training. Additionally, in most of the geometry proofs presented in instructional sessions and textbooks, students are only exposed to the final proof and not the processes involved in performing it (Reiss et al., 2008). Thus, providing process-level PF (Study 1) or metacognitive PF (Study 2) might not seem natural for the preservice mathematics teachers.

As reported in Study 2, different types of PF (e.g., cognitive verification, cognitive elaboration, and self-efficacy PF) were not significantly related to proof comprehension, unlike PF accuracy which was moderately to strongly related to proof comprehension. Furthermore, whereas PF accuracy correlated moderately with CPF beliefs, this was not the case for different types of PF. Additionally, preservice mathematics teachers do not seem to provide specific type(s) of PF based on their understanding of the task, indicating that the more they understand the task does not necessarily mean that they will provide more elaborated PF. Instead, when preservice mathematics teachers have a better understanding of the task, they are only likely to provide more accurate PF. One reason might be because preservice mathematics teachers are not aware of different types or levels of PF (see training effects in Study 1), or that they might not consider them useful.

It can be concluded that preservice mathematics teachers can be trained to provide PF at higher levels (e.g., process-level or self-regulation level), but some levels (Study 1) or types (Study 2) of PF can be more difficult to simulate in the context of geometric tasks. Furthermore, it seems that PF type and PF accuracy convey different information and can be differentially related to PF providers' individual characteristics or outcomes. Therefore, both aspects of PF content (i.e., type and accuracy) should be taken into account by researchers when PF content is under investigation. Whereas having an understanding of the task seems to be associated with more accurate PF, it does not necessarily mean that preservice mathematics teachers will provide more advanced PF types (e.g., cognitive elaboration and process-level or self-regulation level PF). To be able to provide PF with good content preservice mathematics teachers, thus, need to have an adequate knowledge not only of the task but also of different types of PF.

The following sections will address methodological limitations of both studies, followed by theoretical and practical implications. The dissertation will end with recommendations for future research, followed by some general concluding remarks.

4.3. Methodological Limitations

In this dissertation, a quasiexperimental field study was conducted to investigate how PF provision on geometric construction tasks can be implemented within a teacher-training course, followed by an experimental laboratory study to obtain more insight into the processes involved in PF provision. Whereas Study 1 ensured ecological validity, Study 2 provided more control required to explore theoretically proposed processes. Nevertheless, both studies had their methodological limitations.

4.3.1. The challenge of conducting a field experiment

Conducting research in the natural learning environment provides valid measures of learning and improvement; however, it also increases the number of confounding variables. A possible confounding variable in Study 1 is preservice mathematics teachers' adopted learning goals. The participants were informed that their performance in the PF activities would not be assigned an official weight as part of their grade, and they only needed to complete the activities as a part of their course participation. Since we collected their PF and their performance on the basic geometric knowledge test, some students – especially those with low domain knowledge – might have adopted avoidance-oriented goals (i.e., avoiding failure or looking incompetent) and refrained from trying to provide PF at the higher levels.

Another limitation of conducting research within a teacher-training course is the likelihood that preservice teachers from the same track take that course. This on the one hand leads to a more homogeneous sample, yet on the other hand generalization from such research becomes very limited as the findings might not be applicable to all teacher-training classes. In Study 1, many of the preservice mathematics teachers enrolled in the course were majoring in special education, and their training background might have influenced how they responded to the PF training. For instance, one might expect special needs preservice teachers to emphasize praise in their PF. Nevertheless, the proportions of praise provided by participants before and after the training were relatively low compared to other types of PF, as a result of which praise does not appear to be a confounding variable in Study 1.

Conducting research within a course provides the opportunity to implement longitudinal designs to track students' progress throughout the semester. In Study 1, however, preservice mathematics teachers' PF skills were only compared before and after the training.

We did not analyze the progress of the preservice mathematics teachers' PF provision skills during the training sessions in the presence of the instructional scaffolds. One of the challenges to conducting longitudinal research within a course is that not all students attend all the sessions; hence, it is difficult to monitor changes over several sessions. Irregular attendance also resulted in excluding several participants (n = 15) because they did not attend all of the required sessions.

4.3.2. The challenge of conducting eye-tracking research

Although eye-tracking methodology provides promising measures of cognitive processing, a high degree of inference is involved in inferring cognitive processing. In Study 2, eye-tracking measures were accompanied by a performance measure (i.e., proof comprehension) to reduce the degree of inference. However, no relationship was found between the eye-tracking measures and the performance measures, which might be because preservice mathematics teachers also received the peer solution in the second phase when they had to answer the proof comprehension test. This finding highlights the need to accompany eye-tracking measures with cued-retrospective reporting in addition to the performance measures, to reduce the level of inferences and to be able to discover why participants were looking at specific parts of the task (Van Gog et al., 2005; Van Gog & Scheiter, 2010).

Another limitation regarding the eye-tracking measures in Study 2 is that the analyses rely mainly on a single holistic measure, namely proportional total dwell time (PTDT). Although this measure provides information about cognitive processing throughout the learning task, it does not provide information about moment-to-moment processing of different parts of the task (Hyönä, 2010). Using multiple eye-tracking measures, such as (proportional) total dwell time, (proportional) number of fixations, or (proportional) average fixations, and investigating the similarities and differences between these measures would be more informative. Nevertheless, the complexity of calculating fixations from raw data, including choosing the best algorithm and parameters (i.e., dispersion and duration thresholds), remains a challenge and a threat to the accuracy of fixation-based measures. This issue is of particular importance when a slow eye-tracker is used, with a relatively small sample size, which is the case in Study 2. Additionally, longer PTDT could be due to the integration of different types of information (Rayner, 1998). Transitions between the proof's text and figure in the peer solution were not examined in Study 2, although theoretical models of constructing geometry proofs assume a degree of integration between text and figure (e.g., Fischbein, 1993; Ufer et al., 2009). Hence, further analyses can be conducted to examine the two experimental conditions for differences in the proof's text and figure integration in the peer solution.

Eye-movements possess a substantial degree of individual variation (Holmqvist et al., 2011) that can contribute to inflated measurement error. Therefore, a within-subject design is often preferred over a between-subject design. However, in some studies – as in Study 2 – a within-subject design might not be optimal because the first encountered peer solution of the geometry proof is likely to influence how the subsequent peer solution of the same proof would be processed. Another relevant design issue that can contribute to individual variations is the use of self-paced reading techniques. Although this approach is close to the natural reading process, as no time restriction is imposed on participants, it requires them to hit a button to move to the next part. This action can slow down the reading behavior because the reaction time of the finger is slower than the reaction time of the eye (Rayner, 1998). To reduce bias created by individual variations, statistical analyses have been performed only with proportional eye-measures. No extreme outliers were observed in the eye-tracking measures used in Study 2, and non-parametric tests were conducted for variables that violated one of the assumptions of the parametric tests.

4.3.3. Samples

Small sample size was a consistent issue for both studies. In Study 1, the sample size was restricted by class size and the number of preservice mathematics teachers who missed some of the PF training sessions. In Study 2, the labor-intensive data collection using the eye-tracker contributed to the relatively small sample sizes (n = 27 and n = 26). Non-parametric tests were used whenever an assumption of parametric tests was violated.

A more homogeneous sample regarding preservice mathematics teachers' major was acquired in Study 1, whereas the participants in Study 2 had various majors combined with their mathematics teaching major. Thus, preservice mathematics teachers' basic geometric knowledge was used as a covariate in the analysis in Study 2. Equal gender distribution was difficult to acquire, as the majority of students in mathematics teacher education are female. In both studies, females represented the majority of the sample: 79% in Study 1 and 73% in Study 2. In Study 1, random assignment to counter any possible gender effect was not possible, but in Study 2 participants were randomly assigned to conditions. Furthermore, basic geometric knowledge (which is a continuous variable) was used in Study 1 as a grouping variable by dividing the sample into three groups (low, medium, and high domain knowledge). While this is a practical approach to creating groups, it can be a threat to the statistical power (Cohen, 1983).

4.3.4. Measurement of outcome variables

The PF message is one of the main outcomes of the PF provision activity. In this project, different coding schemes were used to analyze the content of the PF message. In Study 1, a coding scheme was adopted based on Hattie and Timperley's (2007) model of feedback. Although, the coding scheme differentiates between PF comment about the self (i.e., self-level or praise), the learning product (i.e., correctness/incorrectness of task performance), learning strategies involved in solving the task (i.e., process-level), and self-regulation strategies (i.e., self-regulation level), it does not make a distinction between verification versus elaboration comments. Additionally, PF often involves comments by the PF providers regarding their incapability to provide PF or to perform the task, and the Hattie and Timperley model does not account for these PF self-efficacy comments. In Study 2, we used a coding scheme developed by Strijbos et al. (2012) that makes a distinction between verification and elaboration comments, and accounts for PF self-efficacy comments. Nevertheless, that coding scheme, despite the fact that it was developed based on PF, does not address inaccuracies in PF providers' comments. The two coding schemes used to analyze PF from Study 1 and Study 2 were adapted for PF provision on geometric tasks. The absence of some types of PF (e.g., metacognitive in Study 2) posits a challenge for the approach of using coding schemes to analyze PF without thoroughly testing the relevance of different types of PF depending on the nature of the learning task, whether it is a geometry proof or a research report. A coding scheme was developed to measure PF accuracy in Study 2, however it only focused on accurate comments. Future studies measuring PF accuracy should also integrate an identification of non-accurate comments in coding schemes to determine PF accuracy more precisely.

In Study 1, PF providers' levels of domain knowledge were determined based Classical Test Theory approach, that especially for the test format we used (i.e., true/false items) does not provide an informative measure or generalizable results regarding of PF providers' domain knowledge because the level of domain knowledge is dependent on the sample. Thus, a low, medium, and high domain knowledge groups might represent different scores depending on different samples used. Instead, using an Item Response Theory approach that takes into account items' difficulties and discrimination serves a better option to measure a trait such as domain knowledge.

The proof comprehension test in Study 2, was used to measure PF providers' comprehension of the geometry proofs they provided PF on. Although the test provided information about some learning outcomes from PF provision, this was only limited to the understanding of the proof at hand and did not include measures of providers' performance on similar tasks. Additionally, while the peer solutions had comparable and non-comparable parts, sub-sections of the proof comprehension test were not sufficiently reliable to allow the comparisons between providers' comprehensions of the comparable and non-comparable parts of the peer solution.

4.4. Theoretical implications

The aim of this dissertation was to investigate PF provision by preservice mathematics teachers on complex geometric tasks that involve scientific reasoning and argumentation by examining: (a) the PF composition process, (b) preservice teachers' individual characteristics (domain knowledge, beliefs about PF provision, PF providers' perceptions of PF, and experienced epistemic emotions), and (c) the relationships between preservice teachers' individual characteristics and the content of provided PF.

Despite the large body of research about PA and PF, there is still no common theoretical process model of PA (Kollar & Fischer, 2010). This issue can be attributed to the lack of empirical studies testing different activities of PA and PF, and different personal and interpersonal characteristics influencing these activities (Strijbos & Sluijsmans, 2010). Furthermore, many studies about PF do not base their research on a clear theoretical framework of PA. In this dissertation, an attempt was made to propose a multi-level multi-process interactive PA framework (i.e., the M^2IPA framework) that includes various PA activities (including PF) and processes and factors likely to influence these activities. Several theoretical perspectives derived from literature reviews on PA (i.e., peer assessment skill model;

Sluijsmans & Prins, 2006), peer assisted learning (Topping & Ehly, 2001), and general feedback literature (i.e., interactional feedback framework; Strijbos & Müller, 2014), were integrated with the purpose of developing a comprehensive theoretical framework of PA activities. Due to the diversity of PA practices and the complexity of different activities (Topping, 1998), it was difficult to validate the entire framework. As such the M^2IPA framework rather served as a basis for the empirical studies reported in this dissertation. The studies focused mainly on one component of the M^2IPA framework, namely PF provision. An attempt was made to empirically examine the PF composition process and several students' individual characteristics likely to influence PF composition and the content of PF (i.e., PF message). Theoretical implications of these findings are discussed next in relation to domain knowledge, training of PA/PF skill, beliefs and perceptions of PA/PF, and the role of epistemic emotions.

Domain knowledge is one of the students' individual characteristics expected to influence the PF provision activity (Patchan & Schunn, 2015). As illustrated in the M^2IPA framework, cognitive conflict is one sub-process induced by the PF activity. The PF provider's and recipient's domain knowledge can trigger cognitive conflict when, for instance, a peer with high domain knowledge provides PF on a solution by low domain knowledge peer and vice versa (Topping, 1998). Although an erroneous peer solution might be expected to trigger more cognitive conflict during PF provision, and hence lead to deeper processing and more error detection and comprehension of the geometry proof, findings of this dissertation showed that PF provided on an erroneous peer solution was less accurate and was associated with less comprehension of the proof. Consistent with proof validation research with high-school students (e.g., Klieme, et al., 2003; Reiss et al., 2000), this dissertation revealed that providing PF on a near-correct peer solution of a geometry proof is easier for preservice mathematics teachers. The value of cognitive conflict lies in that it prompts for engaging in dialogue and students using argumentation to resolve the disagreement (Botvin & Murray, 1975; Johnson & Johnson, 1979). Such dialogue was not the case in both studies since the PF was unidirectional. Cognitive conflict, therefore, should not be automatically assumed as being successfully resolved in each PF activity. More research should be conducted to uncover under which conditions cognitive conflict likely to occur in PF activities leads to deeper understanding of the task and better PF content.

In line with previous research (Gan, 2011; Sluijsmans et al., 2004), Study 1 showed that PF provision skills could be trained. Nonetheless, it also showed that the level of domain knowledge influences how preservice mathematics teachers benefit from the training. It seems that only preservice mathematics teachers with middle and high domain knowledge can provide PF at higher levels after the training. Training PF skills with the support of domain knowledge scaffolds (i.e., worked examples) and PF scaffolds (i.e., prompts and evaluation rubric), cannot compensate for the lack of domain knowledge. Therefore, training PF skills on a specific learning task should follow learning the task and should not be trained concurrently with domain knowledge (Van Zundert, Könings et al., 2012).

Despite the importance of students' beliefs or perceptions in predicting outcomes or behavior, most studies about PF only measured changes in students' beliefs or perceptions of PF (e.g., Sluijsmans et al., 2004; Van Gennip et al., 2010; Wang, 2014). Studies seldom examined the relationship between these individual characteristics and processes and outcomes of the PF activity. This dissertation showed that confidence regarding PF provision is associated with more accurate PF and more certainty by the PF providers that their PF is of high quality, accurate and useful for their peer. Additionally, in line with previous research (e.g., Sluijsmans et al., 2004; Wang, 2014), the PF providers' beliefs about PF provision could change after being involved in several PF activities. PF activities should, therefore, be designed in a way to increase preservice teachers' control over the PF processes, so that they might feel more confident and motivated to provide PF and learn from it.

Study 2 did not only show that some epistemic emotions such as curiosity could be experienced differently depending on the quality of the peer solution, but it also revealed that epistemic emotions such as anxiety and confusion could be associated with less accurate PF. It might be the case that the frequently reported anxiety by students involved in PF (e.g., Cartney, 2010; Harris & Brown, 2013) hinders deeper processing of the learning task during PF activities. Although Study 2 shed some light on the role of emotions, the findings signal the need to investigate further the role that different types of emotions play. These emotions include test emotions likely to be experienced by the PF recipient, epistemic emotions likely to be experienced by the PF provider, and social emotions likely to be experienced by the PF provider and the recipient during PF activities.

Findings in this dissertation regarding PF type (or levels in Study 1) suggest that there might be domain-dependent differences in the type of PF. Particularly, the absence of metacognitive PF in Study 2 – which was observed in a previous study in a different domain (Strijbos et al., 2012) – may indicate that this type of PF is not naturally used when providing PF on geometry proofs. This finding challenges feedback content theories used in PF research, which are usually adopted by research in different domains (e.g., science, psychology, mathematics) with the assumption that different types of PF are equally useful to be used in different domains. Future research should disentangle the role of the domain in shaping the type of PF likely to be considered useful and provided (and used) by students.

To date, there is a lack of theories about PF content. Most PF research adopts models based on findings from teacher's feedback research (e.g., Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Narciss, 2008). The challenges to this approach are that peers are not experts so they might have inaccuracies in their PF, and that they have different roles compared to the teacher (Strijbos & Müller, 2014). Ignoring this issue may result in less effective PF practices. Study 2 revealed different relationships depending on condition (near-correct vs. erroneous peer solution) between PF types, PF accuracy, and students' individual characteristics, which posits challenges for treating PF as teacher feedback. Some researchers argue that the value of PF is in its frequency and immediacy rather than its accuracy (e.g., Ladyshewsky, 2013; Topping, 1998). Empirical evidence, however, suggests that students are more likely to implement PF they receive from more competent peers (Patchan et al., 2013). Furthermore, the findings of this dissertation and previous findings (e.g., Patchan & Schunn, 2015; Van Zundert, Könings et al., 2012) indicate the necessity of domain knowledge for students to be able to provide more accurate PF or PF at higher levels (e.g., self-regulation). That is, PF accuracy is likely to be a prerequisite for higher levels of PF. However, PF accuracy is still widely overlooked in research about PF. These findings indicate the need for a more specific theory about PF content, which should be developed and validated based on PF research.

It was also shown in Study 2 that the mental models for geometry proofs construction, proposed by Ufer et al. (2009), could be applied to geometry proofs validation during PF provision. The model assumes two approaches to proof construction, which were extended for proof validation, namely a figure-based and an analytical approach. Eye-tracking measures revealed that these two approaches could be observed while providing PF on a geometry proof performed by a peer (i.e., validating a proof). However, the quality of the peer solution appears to influence which approach is likely to be adopted by preservice mathematics teachers, with those who are providing PF on an erroneous proof adopting an analytical approach. These findings signal that there is a clear potential for this model to be adopted and tested further in the context of geometry proofs validation.

4.5. Practical implications

This dissertation illustrated that, preservice mathematics teachers' individual characteristics – including domain knowledge (of PF provider and recipient), providers' beliefs about PF provision, providers' perceptions of their PF, and their experienced epistemic emotions – are important for PF provision. Several practical implications can be derived from the findings of this dissertation.

Having adequate levels of domain knowledge and a good understanding of the learning task, are important for preservice mathematics teachers to be able to provide accurate and higher levels of PF (i.e., process and self-regulation). Therefore, it is important to ensure that preservice mathematics teachers can perform the learning tasks (e.g., geometry proofs) on which they receive PF training. Although instructional scaffolds (e.g., PF prompts, worked examples) might work for preservice mathematics teachers with medium and high levels of domain knowledge, they tend to be more challenging for low domain knowledge preservice mathematics teachers. Furthermore, since it is easier for preservice mathematics teachers to validate correct geometry proofs, it is recommended to start PF activities with easy tasks and progressively increase the task difficulty as the preservice mathematics teachers develop adequate domain knowledge and PF provision skills.

Preservice mathematics teachers' confidence regarding PF provision beliefs were observed to be positively related to PF accuracy, whereas anxiety and confusion were negatively related to PF accuracy. Accordingly, it is important to create classroom environments in which students can feel a sense of control over the assessment activities (Cheng & Warren, 1997), experience trust (Carless, 2013), and regard errors as learning opportunities (Borasi, 1994). Additionally, according to the control value theory, anxiety is likely to be experienced when students think that they lack control over their performance

(Pekrun, 2006). Therefore, carefully designing the PF activities so that preservice mathematics teachers perceive control over the activity may help reduce experienced anxiety, for example by involving them in defining assessment criteria (e.g., Sluijsmans et al., 2002) or anonymous PF activities (Vanderhoven et al., 2015) at the start of the PF training.

4.6. Directions for future research

4.6.1. Towards dialogic peer feedback

Peer feedback is in its nature an interactive activity, yet many studies including this dissertation focus only on one of its aspects; provision or reception. Whereas this approach allows disentangling the effects and processes of PF provision and reception, it does not provide information about the full experience. The value of PF activity is likely to be exploited when the feedback loop is closed when the recipient implements the PF (Molloy & Boud, 2013), and when the PF provider gains something out of it through self-reflection and explanation (Nicol et al., 2014). This is unlikely to take place if students do not engage in dialogue and interact during the PF activity. Indeed, recipients do not necessarily apply the PF they receive (e.g., Patchan et al., 2013) which is also the case for teacher feedback (Butler & Winne, 1995). Additionally, PF providers do not always engage in self-reflection while providing PF (Alqassab, Strijbos, & Ufer, 2016). Future research should, therefore, adopt the notion of *dialogic PF*, which Carless (2013) defines as “interactive exchanges in which interpretations are shared, meanings negotiated and expectations clarified” (p. 90). It is through dialogue and interaction that students can seek and provide explanations, justify and negotiate their approaches or judgements, engage in high-level questioning, and engage in argumentation (King, 1994; Kuhn, 2015; Osborne, 2010; Webb, Franke, De, Chan, Freund, Shein, & Melkonian, 2009). Moreover, it was observed in this dissertation that some types of PF were hardly used by preservice mathematics teachers, which was attributed to the nature of the geometric tasks used. Future studies may investigate whether involving preservice mathematics teachers in dialogic PF might foster the provision of metacognitive PF (or process-level and self-regulation level PF) on these types of tasks as in dialogic PF the provider can ask for clarification as to how parts of the proof were performed, or why the recipient adopted one approach instead of another.

4.6.2. Training peer feedback skills: cognitive socio-emotional approach

The role of training PF skills is quite established in the literature, yet the main focus of such training is the content of PF, despite the interpersonal nature of the PF activity. Future studies should consider including socio-emotional aspects in the PF training. This dissertation showed that preservice mathematics teachers' beliefs about PF provision can become less positive in response to PF experiences, and that some of their beliefs and emotions are associated with PF accuracy and their comprehension of the proof. Establishing a supportive environment can make PF activities more constructive. Supportive environment should motivate students to express their ideas, encourage learning from error and acknowledges the role of students as responsible for their own learning and the role of instructors as facilitators and partners in the learning process (Black et al., 2004). In a supportive environment, the role of peers as a source of feedback should be established and students should be allowed to share their insecurities about PF provision and reception to develop a sense of psychological safety (see Edmondson, 1999; Van Gennip et al., 2010). The students should, also, engage in discussions about the purpose of PF as learning activity and the focus of PF on the learning task and not on the self to avoid harm to self-esteem (Molloy et al., 2013) and to develop trust in themselves and in their peers as active learners. Carless (2013) suggested that trust is one interpersonal factor that is likely to facilitate dialogic PF. Empirical findings showed that students' trust improved after being involved in PA intervention (Van Gennip et al., 2010). Additionally, providing structure to the PF activity is suggested to make it more effective

(Ladyshewsky, 2013). Overall, whether socio-emotional support during training can foster more adaptive beliefs, emotions and PF outcomes is still open for future research.

4.6.3. Towards sustainable assessment

The purpose of involving preservice mathematics teachers in PF activities should not only be limited to developing their assessment skills and their domain knowledge to be able to assess and perform a specific task within a course (e.g., geometry proof). PF activities should also prepare learners to be able to meet their own future learning needs, by developing skills that allow them to assess their own and others' learning in the future (Boud & Falchikov, 2006). This is particularly important for preservice teachers as their future career requires them to assess their students' learning, their teaching practices and those of their peers. Thus, research about PF with preservice teachers should aim for sustainable assessment. This is a notion that assessment "encompasses the abilities required to undertake activities that necessarily accompany learning throughout life in formal and informal settings". (Boud, 2000, p. 151). That is, whatever skill students develop by being involved in PF activities should be viable to be employed with similar and different tasks beyond that PF activity. Yet, empirical studies investigating the transfer of PF skills to new contexts are still limited. For this purpose, future research should adopt reflective activities (e.g., dialogic PF), and designs in which PF activities take place over several iterative cycles or throughout the duration of an entire course (i.e., longitudinal design). Additionally, near and far transfer tasks with respect to domain knowledge and PF skills should be used in research about PF to ensure that students develop sustainable learning from being involved in PF activities. For instance, future research may investigate whether students who developed some proof validation skills from being involved in PF provision would then employ these skills to assess their performance while constructing proofs. Research designs for sustainable assessment purposes (e.g., iterative cycles and longitudinal designs) will make it more feasible to empirically test more processes in the

M^{2}IPA framework. Yet, the major challenge to the sustainable assessment approach is to design learning tasks, instructional support, and activities that meet the requirements of sustainable assessment for learners with diverse characteristics and needs.

One of the purposes of sustainable assessment is to contribute to students' confidence to perform an assessment and not undermine it (Boud, 2000). Thus, exploring adaptive and maladaptive emotions and beliefs for PF activities is a central topic for future research. Study 2 was a first attempt in that direction. More research with larger sample sizes is required to examine the relationships between different beliefs and emotions and their impact on PF skills and performance.

4.6.4. Performance and/or assessment skills

For PF skills to be sustainable, assessment should be inseparable of learning (Boud, 2000). Consequently, PF interventions should aim to improve students' domain knowledge and their assessment skills. Future research with preservice teachers should measure their performance on geometric tasks and their PF provision skills, before, during, and after PF interventions. In addition, research about PF should further investigate the relationships between PF types and accuracy with larger sample sizes, and examine the impact of PF types and accuracy on the performance of the PF recipient and the provider. Although, self-reflection is assumed to be a key to learning from PF provision (e.g., Cho & Cho, 2011; Nicol et al., 2014), PF research investigating the role of self-reflection on PF reception and assessee's performance (e.g., Gielen, Peeters, et al., 2010) and PF composition and provision (e.g., Alqassab et al., 2016) is limited. Future research about PF provision should investigate the role of self-reflection prompts on students' learning from providing PF.

4.6.5. Eye-tracking to uncover peer feedback composition

The correlations between PTDT on different components and parts of peer solution and different types of PF (i.e., cognitive verification, cognitive elaboration and self-efficacy PF) and PF accuracy illustrate the potential to employ eye-tracking methodology to explore the PF composition process. Yet, Study 2 was a first attempt to explore the applicability of this methodology to research on PF provision. Eye-tracking can be a useful methodology to provide insights into cognitive processes underlying PF content (type and accuracy) to explore the PF composition process, and also the application of PF by recipients (i.e., PF processing). Therefore, future PF research should use eye-tracking methodology to further explore PF composition, as well as, PF processing to achieve a better understanding of the PF activity. However, research using eye-tracking should rely on accompanying the eye-tracking measures with other type of measures (e.g., cued-retrospective reporting) to reduce the degree of inferences from eye-tracking measures (Van Gog & Scheiter, 2010), especially with relatively unexplored processes such as PF composition. Possible directions for future PF research using eye-tracking methodology include examining PF providers' eye-movements when they use PF provision prompts to compose their PF in relation to PF content, or investigating whether PF training changes the way PF providers look at different parts of the peer solution to provide different types of PF.

4.7. Conclusion

The research presented and discussed in this dissertation has shed some light on the role of preservice mathematics teachers' individual characteristics – including domain knowledge, beliefs and perceptions of PF provision and epistemic emotions – in PF provision on geometric tasks that involve scientific reasoning and argumentation. Overall, the studies suggest that to ensure successful outcomes, preservice mathematics teachers' domain knowledge should be considered when designing and implementing PF activities through, for instance, using individualized instructional support or progressive task difficulty that matches student's domain knowledge. A supportive learning environment should be established to help preservice mathematics teachers to be more confident and to overcome their insecurities regarding PF provision. Overall, educators should strive to make PF feasible for all students, which implies taking students' individual characteristics into account.

References

Ajzen, I. (2005). *Attitudes, personality and behavior* (2nd ed.). England, Open University Press.

Ajzen, I., & Albarracín, D. (2007). Predicting and changing behavior: A reasoned action approach. In I. Ajzen, D. Albarracín, & R. Hornik (Eds.), *Prediction and change of health behavior: Applying the reasoned action approach* (pp. 3–21). London: Lawrence Erlbaum Associates.

Alcock, L., & Weber, K. (2005). Proof validation in real analysis: inferring and checking warrants. *The Journal of Mathematical Behavior, 24*(2), 125–134. doi:10.1016/j.jmathb.2005.03.003.

Alqassab, M., Strijbos, J. W., & Ufer, S. (2016). Exploring the composition process of peer feedback. In Looi, C. K., Polman, J. L., Cress, U., and Reimann, P.(Eds.), *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS), Vol. 2* (pp.862–865). Singapore: International Society of the Learning Sciences.

Balan, A. (2012). *Assessment for learning. A case study in mathematics education* (Doctoral dissertation).Retrieved from: https://dspace.mah.se/bitstream/handle/2043/14356/Assessment%20for%20learning%20%20Andreia%20B.pdf;sequence=2.

Bathke, A. C., Schabenberger, O., Tobias, R. D., & Madden, L. V. (2009). Greenhouse-Geisser adjustment and the ANOVA-Type statistic: cousins or twins? *The American Statistician, 63*(3), 239–246. doi:10.1198/tast.2009.08187

Bednarik, R., & Tukiainen, M. (2008). Temporal eye-tracking data: Evolution of debugging strategies with multiple representations. In K. J. Rähä & A. Duchowski (Eds.), *Proceedings of the 2008 symposium on eye tracking research & applications* (pp. 99–102). New York, NY: ACM. doi:10.1145/1344471.1344497

Beitlich, J. T., Obersteiner, A., & Reiss, K. (2015). How do secondary school students make use of different representation formats in heuristic worked examples? An analysis of eye movements. In In K. Beswick, T. Muir, & J. Wells (Eds.), *Proceedings of the 39th Conference of the International Group for the Psychology of Mathematics Education, Vol. 2* (pp. 97–104). Hobart, Australia: PME.

Bernichon, T., Cook, K. E., & Brown, J. D. (2003). Seeking self-evaluative feedback: The interactive role of global self-esteem and specific self-views. *Journal of Personality and Social Psychology, 84*(1), 194–204. doi:10.1037/0022-3514.84.1.194

Black, H. (1986). Assessment for learning. In D. L. Nuttall (Ed.), *Assessing educational achievement* (pp. 7–18). London, England: Falmer Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). *Inside the black box*. London, England: School of Education, King's College London.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(5), 5–31. doi:10.1007/s11092-008-9068-5

Boero, P. (1999). Argumentation and mathematical proof: a complex, productive, unavoidable relationship in mathematics and mathematics education. *International Newsletter on the Teaching and Learning of Mathematical Proof, 7*(8), Retrieved from:http://www-didactique.imag.fr/preuve/.

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

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education*, *25*(2), 166–208. doi:10.2307/749507

Botvin, G. J., & Murray, F. B. (1975). The efficacy of peer modeling and social conflict in the acquisition of conservation. *Child Development*, *46*(3), 796–799. doi:10.2307/1128583

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151–167. doi: 10.1080/713695728

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399–413. doi:10.1080/02602930600679050

Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.

Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education*, *28*(7), 968–978. doi:10.1016/j.tate.2012.05.003

Brown, G. T. L., Peterson, E. R., & Yao, E. S. (2016). Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. *British Journal of Educational Psychology*. Advance online publication. doi: 10.1111/bjep.12126

Brunner, E., Munzel, U., Puri, M. L. (1999). Rank score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis, 70*(2), 286–317. doi:10.1006/jmva.1999.1821

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. doi:10.3102/00346543065003245

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction, 20*, 155–166. doi:10.1016/j.learninstruc.2009.02.014

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233. doi:10.1080/03075070600572132

Carless, D. (2013). Trust and its role in facilitating dialogic feedback. In D. Boud & L. Molloy (Eds.), *Effective feedback in higher and professional education* (pp. 90–103). London: Routledge.

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*(3), 360–367. doi:10.1037/0022-0663.86.3.360

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education, 35*(5), 551–564. doi:10.1080/02602931003632381

Chen, N., Wei, C., Wua, K., & Uden, L. (2009). Effects of high level prompts and peer assessment on online students' reflection levels. *Computers & Education, 52*(2), 283–291. doi:10.1016/j.compedu.2008.08.007

Cheng, K.H., Hou, H. T., & Wu, S. Y. (2014). Exploring students' emotional responses and participation in an online peer assessment activity: A case study. *Interactive Learning Environments, 22*(3), 271–287. doi:10.1080/10494820.2011.649766

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

Chinn, K. A., & Clark, D. B. (2013). Learning through collaborative argumentation. In C. E. Hmelo-Silver, C. A. Chinn, C. K. K. Chan, & A. M. O'Donnel (Eds.),*The international handbook of collaborative learning* (pp. 315–332).New York, NY: Routledge.

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

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

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

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. doi:10.1037/0022-0663.98.4.891

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249–253.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. New Jersey, NJ: Lawrence Erlbaum Associates.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. doi:10.1016/j.learninstruc.2011.10.001

Dreyfus, T. (1999). Why Johnny can't prove. *Educational Studies in Mathematics, 38*(1), 85–109. doi:10.1023/A:1003660018579

Dunbar, K., & Fugelsang, J. (2004). Scientific thinking and reasoning. In K. J Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705–726). New York, NY: Cambridge University Press.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383.

Edwards, A. R., Esmonde, I., & Wagner, J. F. (2011). Learning mathematics. In R. E. Mayer & P. A. alexander (Eds.), *Handbook of research on learning and instruction* (pp. 55–77). New York, NY: Routledge.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725–747. doi:10.1037/0003-066X.49.8.725

Ericsson, K. A., & Crutcher, R. J. (1991). Introspection and verbal reports on cognitive processes—two approaches to the study of thinking: A response to Howe*. *New Ideas in Psychology, 9*(1), 57–91. doi:10.1016/0732-118X(91)90041-J

Ernst, D. C., Hodge. A., & Schultz, A. (2015). Enhancing proof writing via cross-institutional peer review. *PRIMUS: Problems, Resources, and Issues in Mathematics and Undergraduate Studies, 25*(2), 121–130, doi:10.1080/10511970.2014.921652

Ertmer, P. A., Richardson, J. C., Lehman, J. D., Newby, T. J., Cheng, X., Mong, C., & Sadaf, A. (2010). Peer feedback in a large undergraduate blended course: Perceptions of value and learning. *Journal of Educational Computing Research*, *43*(1), 67–88. doi:10.2190/EC.43.1.e

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. doi:10.3102/0034654312474350

Falchikov, N., & Boud, D.(2007). Assessment and emotion: the impact of being assessed. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for longer term* (pp. 144–158). London: Routledge.

Falchikov, N., & Godlfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. doi: 10.3102/00346543070003287

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed). Los Angeles: SAGE.

Fischbein, E. (1993). The theory of figural concepts. *Educational Studies in Mathematics, 24*(2), 139–162.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer., S., Fischer, M., Strijbos, J. W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45. doi:10.14786/flr.v2i2.96

Fodor, E. M., & Carver, R. A. (2000). Achievement and power motives, performance feedback, and creativity. *Journal of Research in Personality, 34*(4), 380–396. doi:10.1006/jrpe.2000.2289

Gallimore, M., & Stewart, J. (2014). Increasing the impact of mathematics support on aiding student transition in higher education. *Teaching Mathematics and Its Applications, 33*(2), 98–109. doi:10.1093/teamat/hru008

Gan, M. (2011). *The effects of prompts and explicit coaching on peer feedback quality* (Unpublished doctoral dissertation). The University of Auckland, New Zealand.

Gan, M. J. S., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science, 42*(6), 861–878. doi:10.1007/s11251-014-9319-4

Ghazan, D. (1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics, 24*(4), 359–387.

Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning*, *31*(5), 435–449. doi:10.1111/jcal.12096

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

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

Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback and of various peer feedback forms in a secondary school writing curriculum. *British Educational Research Journal*, *36*(1), 143–162. doi:10.1080/01411920902894070

Graesser, A. C., D'Mello, S., & Cade, W. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 406–426). New York, NY: Taylor & Francis.

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. doi:10.1016/j.learninstruc.2007.09.008

Guasch, T., Espasa, A., Alvarez, I. M., & Kirschner, P. A. (2013). Effects of feedback on collaborative writing in an online learning environment. *Distance Education*, *34*(3), 324–338. doi:10.1080/01587919.2013.835772

Gukas, I. D., Miles, S., Heylings, D. J., & Leinster, S. J. (2008). Medical students' perceptions of peer feedback on an anatomy student-selected study module. *Medical Teacher*, *30*(8), 812–814. doi:10.1080/01421590802262070

Gulwani, S., Korthikanti, V. A., & Tiwari, A. (2011). Synthesizing geometry constructions. In M. Hall & D. Padu (Eds.), *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation* (pp. 50–61), New York: ACM. doi:10.1145/1993498.1993505

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education*, *36,* 101–111. doi:10.1016/j.tate.2013.07.008

Harris, L. R., Brown, G. T. L., & Harnett, J. (2015). Analysis of New Zealand primary and secondary student peer- and self-assessment comments: Applying Hattie & Timperley's feedback model. *Assessment in Education: Principles, Policy and Practice*, *22*(2), 265–281.doi:10.1080/0969594X.2014.976541

Hattie, J. A., & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York, NY: Routledge.

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

Healy, L.; Hoyles, C. (1998): *Justifying and proving in school mathematics. Technical report on the nationwide survey*. Institute of Education, University of London.

Heemsoth, T., & Heinze, A. (2014). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science*, *42*(4), 639–657. doi:10.1007/s11251-013-9302-5

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology, 84*(1), 76–84. doi:10.1037/0022-0663.84.1.76

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18–32. doi:10.1037/0022-0663.87.1.18

Heinze, A., & Reiss, K. (2001). Reasoning and proof in the mathematics classroom. *Analysis*, *27*(2–3), 331–366.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121. doi:10.1177/109442819800100106

Hodds, M., Alcock, L., & Inglis, M. (2014) Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*(1), 62–101. doi:10.5951/jresematheduc.45.1.0062

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & De Weijer, J. V. (2011). *Eye tracking: A comprehensive guide to methods and measures*. New York: Oxford University press.

Hou, H.T., & Cheng, K.H. (2012). Analyzing the latent emotional transfer pattern (LETP) of a learning community in an online peer-assessment activity: Colloquium. *British Journal of Educational Technology*, *43*(4), E113–E116. doi: 10.1111/j.1467-8535.2012.01301.x

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, *71*, 133–152. doi: 10.1016/j.compedu.2013.09.019

Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson Wadsworth.

Hu, C., Wang, Q., Fu, G., Quinn, P. C., & Lee, K. (2014). Both children and adults scan faces of own and other races differently. *Vision Research*, *102*, 1–10. doi:10.1016/j.visres.2014.05.010

Hyönä, J. (2010). The use of eye-movements in the study of multimedia learning. *Learning and Instruction*, *20*(2), 172–176. doi:10.1016/j.learninstruc.2009.02.013

Ilgen, D. R., & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology: An International Review*, 49(3), 550–565. doi:10.1111/1464-0597.00031

Illich, I. (1971). *Deschooling society*. Manchester, UK: Pelican Books.

Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, *43*(4), 358–390. doi:10.5951/jresematheduc.43.4.0358

Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., & McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? In C. Delgado Kloos, D. Gillet, R. M. Crespo Gracia, F. Wild, & M. Wolpers (Eds.), 6th *European Conference on Technology Enhanced Learning. Towards Ubiquitous Learning* (pp. 181–195). Palermo, Italy: Springer.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Johnson, D. W., & Johnson, R. T. (1979). Conflict in the classroom: Controversy and learning. *Review of Educational Research*, *49*(1), 51. doi:10.2307/1169926

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge: Harvard University Press.

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418–439.

Jolly; B., & Boud, D. (2013). Written feedback: What is it good for and how can we do it well? In D. Boud & E. Molloy (Eds.), *Feedback in higher education and professional education: Understanding it and doing it well* (pp. 104–124). New York, NY: Routledge.

Jones, K. (2000). Providing a foundation for deductive reasoning: Students' interpretations when using dynamic geometry software and their evolving mathematical explanations. *Educational Studies in Mathematics, 44* (1), 55–85. doi:10.1023/A:1012789201736

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787. doi: 10.1080/03075079.2013.821974

Jonsson, A. (2012). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, *14*(1), 63–76. doi:10.1177/1469787412467125

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processing. *Cognitive Psychology, 8*(4), 441–480. doi: 10.1016/0010-0285(76)90015-3

Just, M. A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. doi:10.1037/0033-295X.87.4.329

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20*(8), 963–973.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*(2), 338–368. doi:10.3102/00028312031002338

King, A. (2002). Structuring peer interaction to promote high-level cognitive processing. *Theory Into Practice, 41*(1), 33–39. doi:10.1207/s15430421tip4101_6

King, P. E., Schrodt, P., & Weisel, J. J. (2009). The instructional feedback orientation scale: conceptualizing and validating a new measure for assessing perceptions of instructional feedback. *Communication Education*, *58*(2), 235–261. doi: 10.1080/03634520802515705

Klenowski, V. (2009). Editorial: Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy, and Practice, 16*(3), 263–268.

Klieme, E., Reiss, K. & Heinze, A. (2003). Geometrical competence and understanding of proof: A study based on TIMSS items. In F.L. Lin & J. Guo (Eds), *Proceedings of the International Conference on Science and Mathematics Learning 2003* (pp. 60–80). Taipei, Taiwan: National Taiwan Normal University.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284. doi:10.1037/0033-2909.119.2.254

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, *14*(4), 511–550.

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*(4), 344–348. doi:10.1016/j.learninstruc.2009.08.005

Kuhn, D. (2010). What is scientific thinking and how does it develop? In U Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 497–519). United Kingdom, UK: Blackwell Publishing.

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher, 44*(1), 46–53 doi:10.3102/0013189X15569530

Kuzniak, A., & Houdement, C. (2001). Pretty (good) didactical provocation as a tool for teacher's training in geometry. In J. Novotná (Ed.), *Proceedings of European Research in Mathematics Education II* (pp. 292–303). Mariánské Lázně Czech Republic: Charles University, Faculty of Education.

Ladyshewsky, R. K. (2013). The role of peers in feedback processes. In D. Boud & E. Molloy (Eds.), *Feedback in higher education and professional education: Understanding it and doing it well* (pp. 174–189). New York, NY: Routledge.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–10. doi:10.3389/fpsyg.2013.00863

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100.

Lauf, L., & Dole, S. (2010). Assessment for learning tasks and the peer assessment process. In L. Sparrow, B. Kissane, & C. Hurst (Eds.). *Shaping the future of mathematics education. Proceedings of the 33rd annual conference of the Mathematics Education Research Group of Australasia* (pp. 320–327). Fremantle, Australia: MERGA.

Lavy, I., & Shriki., A. (2014). Engaging prospective teachers in peer assessment as both assessors and assesses: The case of geometrical proofs. *International Journal for Mathematics Teaching and Learning.* Retrieved from http://www.cimt.plymouth.ac.uk/jour-nal/lavey2pdf

Li, H., Xiong, Y., Zang, X., Kronhaber, M. L., Lyu, Y., Chung. K. S., & Suen., H. K. (2015). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. doi:10.1080/02602938.2014.999746

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

Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the feedback orientation scale (FOS). *Journal of Management*, *36*(6), 1372–1405. doi:10.1177/0149206310373145

Lingefjärd, T., & Holmquist, M. (2005). To assess students' attitudes, skills and competencies in mathematical modeling. *Teaching Mathematics and its Applications*, *24*(2-3), 123–133. doi:10.1093/teamat/hri021

Mair, P., & Wilcox, R. (2015). Robust statistical methods: The R package WRS2. Retrieved from http://cran.usthb.dz/web/packages/WRS2/vignettes/WRS2.pdf

Manin, A. (2010). *A course in mathematical logic for mathematicians*. New York, NY: Springer.

Mason, B. J., & Brnning, R.H. (2001). *Providing feedback in computer-based instruction: What the research tells us* (CLASS Research Report No. 9). Center for Instructional Innovation, University of Nebraska-Lincoln.

Meyer, K., Rasch, T., & Schnotz, W. (2010). Effects of animation's speed of presentation on perceptual processing and learning. *Learning and Instruction*, *20*(2), 136–145. doi: 10.1016/j.learninstruc.2009.02.016

Miller, L. M., & West, R. L. (2010). The effects of age, control beliefs, and feedback on self-regulation of reading and problem solving. *Experimental Aging Research: An International Journal Devoted to Scientific Study of the Aging Process*, *36*(1), 40–63. doi:10.1080/03610730903418380

Miyakawa, T. (2004). Reflective symmetry in construction and proving. In M. J. Høines & A. B. Berit (Eds.), *Proceedings of the 28th Conference of the International Group for the*

Psychology of Mathematics Education (pp. 337–344). Bergen, Norway: Bergen University College.

Molloy, E., & Boud, D. (2013). Changing conceptions of feedback. In D. Boud & E. Molloy (Eds.), *Feedback in higher education and professional education: Understanding it and doing it well* (pp. 12–33). New York, NY: Routledge.

Molloy, E., Borrell-Carrió, F., & Epstein, R. (2013). The impact of emotions in feedback. In D. Boud & E. Molloy (Eds.), *Feedback in higher education and professional education: Understanding it and doing it well* (pp. 50–71). New York, NY: Routledge.

Muis, K. R., Psaradellis, C., Lajoie, S. P., Di Leo, I., & Chevrier, M. (2015). The role of epistemic emotions in mathematics problem solving. *Contemporary Educational Psychology*, *42*, 172–185. doi: 10.1016/j.cedpsych.2015.06.003

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. (Pellegrino, J. W. & Hilton, M. L., Eds.) Washington, DC: The National Academies Press.

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. 125–144). Mahwah, NJ: Lawrence Erlbaum.

Narciss, S. (in press). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In D. Carless, S. Bridges, C.K.W. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. XX–XX). Dordrecht: Springer.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In in H. M. Niegemann, D. Leutner, & R. Brünken (Ed.).*Instructional Design for Multimedia Learning* (pp. 181–195). Münster, Germany: Waxmann.

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

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12), 1–23. doi:10.18637/jss.v050.i12

Nückles, M., Hübner, S., Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271. doi:10.1016/j.learninstruc.2008.05.002

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 396–402.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way anova. *Behaviormetrika, 40*(2), 129–147. doi:10.2333/bhmk.40.129

Okita, S. Y. (2014). Learning from the folly of others: Learning to self-correct by monitoring the reasoning of virtual characters in a computer-supported mathematics learning environment. *Computers & Education, 71*, 257–278. doi:10.1016/j.compedu.2013.09.018

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*, 463–466. doi:10.1126/science.1183944

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122–133.

Panadero, E., Brown, G. T. L., & Strijbos, J. W. (in press). The future of student self-assessment: A review of known unknown and potential directions. *Educational Psychology Review*.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005

Patchan, M. M., Charney, D., & Schunn, C. D. (2009). A validation study of students' end comments: Comparing comments by students, a writing instructor, and a content instructor. *Journal of Writing Research*, *1*(2), 124–152. doi:10.17239/jowr-2008.01.01.1

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

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science, 43*(5), 591–614. doi:10.1007/s11251-015-9353-x

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. doi:10.1007/s10648-006-9029-9

Pekrun, R. (2011). Emotions as drivers of learning and cognitive development. In R. A. Calvo & S. K. D'Mello (Eds.), *New Perspectives on Affect and Learning Technologies* (pp. 23–39). New York, NY: Springer.

Pekrun, R., Cusack, A., Murayama, K., Elliot, A. J., & Thomas, K. (2014). The power of anticipated feedback: Effects on students' achievement goals and achievement emotions. *Learning and Instruction*, *29*, 115–124. doi:10.1016/j.learninstruc.2013.09.002

Pekrun, R., & Meier, E. (2011). *Epistemic emotion scales (EES)*. Unpublished manuscript, Department of Psychology, University of Munich, Munich, Germany.

Pekrun, R., & Stephens, E. J. (2012). Academic emotions. In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (pp. 3–31). Washington: American Psychological Association.

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2016). Measuring emotions during epistemic activities: The epistemically-related emotion scales. *Cognition and Emotion,* 1–9. doi:10.1080/02699931.2016.1204989

Piaget, J. (1959). *The language and thought of the child* (3rd ed.). New York, NY: Routledge.

Praver, M., Rouault, G., & Eidswick, J. (2011). Attitudes and affect toward peer evaluation in EFL reading circles. *Reading*, *11*(2). Retrieved from http://www.readingmatrix.com/articles/april_2011/praver_rouault_eidswick.pdf

Price, M., Handley. K., Millar. J., & O'Donovan, B (2010). Feedback: All the effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*(3), 277–289. doi:10.1080/02602930903541007

Prins, F. J., Sluijsmans, D. M. A., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education*, *11*(3), 289–303. doi: 10.1007/s10459-005-3250-z

Raes, A., Vanderhoven, E., & Schellens, T. (2013). Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. *Studies in Higher Education*, *40*(1), 178–193. doi: 10.1080/03075079.2013.823930

Raney, G. E., & Rayner, K. (1995). Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology*, *49*(2), 151–172.

Rayner, K. (1998). Eye-movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. doi:10.1037//0033-2909.124.3.372

Rayner, K. (2009). Eye-movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506. doi:10.1080/17470210902816461

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.

Reiss, K., & Heinze, A. (2001). Knowledge acquisition in students' argumentation and proof processes. In G. Törner, R. Bruder, N. Neill, A. Peter-Koop & B. Wollring (Eds.), *Developments in Mathematics Education in German-Speaking Countries. Selected Papers from the Annual Conference on Didactics of Mathematics* (pp. 107–115). Hildesheim: Franzbecker.

Reiss, K., Heinze, A. (2004). The teaching of proof at the lower secondary level- a video study. *ZDM Mathematics Education, 36*(3), 98–104. doi: 10.1007/BF02652777

Reiss, K., Heinze, A. & Klieme, E. (2000). Argumentation, proof and the understanding of proof. In H.-G. Weigand, N. Neill, A. Peter-Koop, K. Reiss, G. Törner, & B. Wollring (Eds.), Developments in mathematics education in Germany: *Selected papers from the* *annual Conference on Didactics of Mathematics* (pp. 109–120). Hildesheim, Germany: Franzbecker.

Reiss, K. M., Heinze, A., Renkl, A., & Groß, C. (2008). Reasoning and proof in geometry: Effects of a learning environment based on heuristic worked-out examples. *ZDM Mathematics Education*, *40*(3), 455–467. doi: 10.1007/s11858-008-0105-0

Reiss, K. Hellmich, F., & Reiss, M. (2002). Reasoning and proof in geometry: Prerequisites of knowledge acquisition in secondary school students. In A. Crockburn & E. Nardi (Eds.), *Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education Vol. 3* (pp.113–120). Norwich: University of East Anglia.

Reiss, K.; Klieme, E.; Heinze, A. (2001): Prerequisites for the understanding of proofs in the geometry classroom. In M.Van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education Vol. 4.*(pp.97–104).Utrecht: Utrecht University.

Renkl, A. (2011). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 272–295). New York, NY: Routledge.

Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. *Handbook of Research on Teacher Education*, *2*, 102–119.

Rubie-Davies, C. M., Flint, A., & McDonald, L. G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships?: Teacher efficacy, expectations, and goal orientation. *British Journal of Educational Psychology*, *82*(2), 270–288. doi:10.1111/j.2044-8279.2011.02025.x

Sadler, D. R. (1989). Formative assessment and the design of instructional system. *Instructional Science*, *18*(2), 119–144. doi:10.1007/BF00117714

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

Santos, L. & Pinto, L. (2009). Lights and shadows of feedback in mathematics learning. 2009. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Eds.). *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education, Vol. 5* (pp. 49–56). Thessaloniki, Greece: PME

Schoenfeld, A. H. (1986). On having and using geometric knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 225–264). Hillsdale, NJ: Erlbaum.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well taught" mathematics courses. *Educational Psychologist, 23*(2), 145–166.

Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for Research in Mathematics Education, 20*(4), 338–355. doi: 10.2307/749440

Selden, A., & Selden, J. (2003). Validations of proofs considered as texts: can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education, 34*(1), 4–36. doi:10.2307/30034698

Selden, A., & Selden, J. (2015a, December). *A comparison of proof comprehension, proof construction, proof validation and proof evaluation*. Paper presented at the KHDM Conference Didactics of Mathematics in Higher Education as a Scientific Discipline, Hannover, Germany.

Selden, A., & Selden, J. (2015b). Validations of proofs as a type of reading and sense-making. In In K. Beswick, T. Muir, & J. Wells (Eds.), *Proceedings of the 39th Conference of the International Group for the Psychology of Mathematics Education*: *Vol. 4.* (pp. 145–152).Hobart, Tasmania.

Siegler, R.S. (2002) Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment, transition processes in development and learning* (pp. 31–58). Cambridge: Cambridge University Press.

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

Sluijsmans, D. M. A., Brand-Gruwel, S., & Van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454. doi:0.1080/0260293022000009311

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

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. Studies in Educational Evaluation, 32, 6–22. doi: 10.1016/j.stueduc.2006.01.005

Smith, A. D. (2001). Perception and belief. *Philosophy and Phenomenological Research*, *62*(2), 283–309.

Sommerhoff, D., Ufer, S., Kollar, I. (2016). Proof validation aspects and cognitive student prerequisites in undergraduate mathematics. In C. Csíkos, A. Rausch & J. Szitányi, (Eds.), *Proceedings of the 40th Conference of the International Group for the Psychology of Mathematics Education*, *Vol. 4.* (pp. 219–226). Szeged, Hungary: PME.

Strijbos, J. W., Martens, R., L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computers & Education, 46*(1), 29–48. doi:10.1016/j.compedu.2005.04.002

Strijbos, J. W., & Müller, A. (2014). Personale faktoren im feedbackprozess. In H. Ditton & A. Müller (Eds.), *Feedback und rückmeldungen: Theoretische grundlagen, empirische befunde, praktische anwendungsfelder* [Feedback and evaluation: Theoretical foundations, empirical findings, practical implementation] (pp. 87–134). Münster, Germany: Waxmann.

Strijbos, J. W., Narciss, S., Dünnebier, K. (2010). Peer feedback content and sender's comptence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303. doi: 10.1016/j.learninstruc.2009.08.008

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., & Segers, M. S. R. (2009). Fostering interactivitiy through formative peer assessment in (web-based) collaborative learning environments. In M. Constantinos, N. Tsianos, & P. Germanakos (Eds.), *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 375–395). New York, NY: Information Science Reference.

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

Strijbos, J. W., Van Goozen, B., & Prins, F. (2012, August). *Developing a coding scheme for analysing peer feedback messages*. Paper presented at the EARLI-SIG 1 Assessment and Evaluation Conference, Brussels, Belgium.

Suppes, P., Cohen, M., Laddaga, R., Anliker, J., & Floyd, R. (1983). A procedural theory of eye-movements in doing arithmetic. *Journal of Mathematical Psychology*, *27*(4), 341–369. doi:10.1016/0022-2496(83)90033-0

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Essex, England: Pearson.

Tapan, M., S., & Arslan, C. (2009). Preservice teachers' use of spatio-visual elements and their level of justification dealing with a geometrical construction problem. *US-China Education Review, 6*(3), 54–60.

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

Topping, K. J., & Ehly, S. W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*(2), 113–132. doi:10.1207/S1532768XJEPC1202_03

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

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on phrasing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 3*(3), 285–318. doi: 10.1006/jmla.1994.1014

Tsai, M., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education, 58*(1), 375–385.doi:10.1016/j.compedu.2011.07.012

Tsovaltzi, D., Melis, E., McLaren, B. M., Meyer, A.K., Dietrich, M., & Goguadze, G. (2010). Learning from erroneous examples: When and how do students benefit from them? In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Sustaining TEL: From Innovation to Learning and Practice, Vol.3683* (pp. 357–373). Berlin/Heidelberg: Springer.

Ufer, S., Heinze, A., Reiss, K. (2008). Individual predictors of geometrical proof competence. In Figueras, O., Cortina, J.L., Alatorre, S., Rojano, T., Sepulveda, A. (Eds.), *Proceedings of the Joint Meeting of PME 32 and PME-NA XXX* (Vol. 4, pp. 361-368). Mexico: Cinvestav-UMSNH.

Ufer, S., Heinze, A., & Reiss, K. (2009). Mental models and the development of geometric proof competency. In M. Tzekaki, M. Kaldrimidou & C. Sakonidis (Eds.). *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education, Vol. 5* (pp. 257-264). Thessaloniki, Greece: PME.

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

Van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19–36. doi:10.1080/02602930500262346

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes A meta-analysis. *Review of Educational Research, 85*(4), 475–511. doi:10.3102/0034654314564881

VandeWalle, D., & Cummings, L. L. (1997). A test of the influence of goal orientation on the feedback-seeking process. *Journal of Applied Psychology, 82*(3), 390–400.

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*, 41-554. doi:10.1016/j.edurev.2008.11.002

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280–290. doi: 10.1016/j.learninstruc.2009.08.010

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data: Uncovering expertise-related differences. *Applied Cognitive Psychology*, *19*(2), 205–221. doi:10.1002/acp.1112

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*(2), 95–99. doi:10.1016/j.learninstruc.2009.02.009

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction*, *20*(4), 316–327. doi:10.1016/j.learninstruc.2009.08.009

Van Zundert, M., J., Könings, K. D., Sluijsmans, D., M. A., & Van Merriënboer, J. J. G. (2012). Teaching domain-specific skills before peer assessment skills is superior to teaching them simultaneously. *Educational Studies*, *38*(5), 541–557. doi:10.1080/03055698.2012.654920

Van Zundert, M., J., Sluijsmans, D., M. A., Könings, K. D., & Van Merriënboer, J. J. G. (2012). The differential effects of task complexity on domain-specific and peer assessment skills. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *32*(1), 127–145. doi:10.1080/01443410.2011.626122

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

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasiexperimental study. *Computers & Education*, *81*, 123–132. doi: 10.1016/j.compedu.2014.10.001

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman., Eds.) (A. R. Luria, M. Lopez-Morillas & M. Cole [with J. V. Wertsch], Trans.) Cambridge, Mass.: Harvard University Press. (Original manuscripts [ca. 1930-1934])

Walker, M. (2015). The quality of written peer feedback on undergraduates' draft answers to an assignment, and the use made of the feedback. *Assessment & Evaluation in Higher Education*, *40*(2), 232–247. doi: 10.1080/02602938.2014.898737

Wang, W. (2014). Students' perceptions of rubric-referenced peer feedback on EFL writing: A longitudinal inquiry. *Assessing Writing, 19*, 80–96. doi: 10.1016/j.asw.2013.11.008

Webb, N. M., Franke, M. L., De, T., Chan, A. G., Freund, D., Shein, P., & Melkonian, D. K. (2009). "Explain to your partner": teachers' instructional practices and students' dialogue in small groups. *Cambridge Journal of Education*, *39*(1), 49–70. doi:10.1080/03057640802701986

Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics*, *56*(2–3), 209–234. doi:10.1023/B:EDUC.0000040410.57253.a1

Weiner, B. (2007). Examining emotional diversity in the classroom: An attribution theorist considers the moral emotions. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 75–88). Burlington, MA: Elsevier.

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Oxford, England: Elsevier.

Wilcox, R. R. (2013). A heteroscedastic method for comparing regression lines at specified design points when using a robust regression estimator. *Journal of Data Science: JDS*, *11*(2), 281–291.

Wiliam, D. (1999a). Formative assessment in mathematics part 1: Rich questioning. *Equals: Mathematics and Special Educational Needs*, 5(2), 15–18.

Wiliam, D. (1999b). Formative assessment in mathematics part 2: Feedback. *Equals: Mathematics and Special Educational Needs*, *5*(3), 8–11.

Wiliam, D. (2000). Formative assessment in mathematics part 3: The learners' role. *Equals: Mathematics and Special Educational Needs*, *6*(1), 19–22.

Wiliam, D., & Thompson. M. (2007). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.

Williams, J. R., & Johnson, M. A. (2000). Self-supervisor agreement: The influence of feedback seeking on the relationship between self and supervisor ratings of performance. *Journal of Applied Social Psychology*, *30*(2), 275–292. doi: 10.1111/j.1559-1816.2000.tb02316.x

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education,27*(4), 458–477.

Yang, K. L., & Lin, F. L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*(1), 59–76. doi:10.1007/s10649-007-9080-6

Yi, L., Liu, Y., Li, Y., Fan, Y., Huang, D., & Gao, D. (2012). Visual scanning patterns during the dimensional change card sorting task in children with autism spectrum disorder. *Autism Research and Treatment*, *2012*, 1–11. doi:10.1155/2012/123053

Zerr, J. M., & Zerr, R. J. (2011). Learning from their mistakes: Using students' incorrect proofs as a pedagogical tool. *PRIMUS, 21*(6), 530–544. doi:10.1080/105119709033869