University of Groningen

The impact of peer solution quality on peer-feedback provision on geometry proofs

Alqassab, Maryam; Strijbos, Jan-Willem; Ufer, Stefan

Published in:
Learning and Instruction

DOI:
[10.1016/j.learninstruc.2018.07.003](https://doi.org/10.1016/j.learninstruc.2018.07.003)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Alqassab, M., Strijbos, J-W., & Ufer, S. (2018). The impact of peer solution quality on peer-feedback provision on geometry proofs: Evidence from eye-movement analysis. *Learning and Instruction*, 58, 182-192. <https://doi.org/10.1016/j.learninstruc.2018.07.003>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

The impact of peer solution quality on peer-feedback provision on geometry proofs: Evidence from eye-movement analysis

Maryam Alqassab^{a,*}, Jan-Willem Strijbos^{a,b}, Stefan Ufer^c

^a Department of Psychology, LMU Munich, Germany

^b Faculty of Behavioural and Social Sciences, University of Groningen, the Netherlands

^c Mathematics Institute, LMU Munich, Germany

ARTICLE INFO

Keywords:

Peer-feedback
Geometry proofs
Eye-tracking
Preservice teachers

ABSTRACT

Providing feedback on peer solutions to geometry proofs can support preservice mathematics teachers' assessment skills of such complex tasks. However, the quality of peer solutions may influence cognitive processing during peer-feedback provision, learning from providing peer-feedback, and peer-feedback content. To investigate this effect, we recorded the eye-movements of fifty-three preservice mathematics teachers while providing feedback on a near-correct or an erroneous peer solution to a geometry proof, and we measured their proof comprehension and peer-feedback content. Results show that the absence of errors earlier in the peer solution facilitated reliance on a figure-based approach, whereas encountering errors earlier in the peer solution was associated with more focus on the text of the proof. Students who provided peer-feedback on the near-correct peer solution had better comprehension of the proof, and they provided more accurate peer-feedback. Errors in peer solutions thus appear to hinder positive peer-feedback outcomes.

1. Introduction

Proof is central to mathematics instruction in schools (National Council of Teachers of Mathematics, 2000). Yet, there is a consensus in the literature that proof is challenging for high-school and university students (for a review see Harel & Sowder, 2007). In this context, proof is understood as a more or less formal, deductive argumentation establishing the validity of a mathematical statement, based on definitions and proven theorems from a framework theory (Stylianides, 2007). Students' weakness in proof is attributed to several factors including passive learning (e.g., observing a teacher, learning from textbooks). Research has shown that making proof instruction more active (e.g., through self-explanation) can improve students' understanding of proofs (i.e., proof comprehension; Hodds, Alcock, & Inglis, 2014).

Peer-feedback has the potential to stimulate active learning of proofs. Providing peer-feedback involves judging the correctness of a peer solution (e.g., proof) and producing statements to support or explain these judgements. Unlike other active learning techniques (e.g., self-explanation), peer-feedback on proofs involves judging the correctness of a proof constructed by another source (i.e., the peer). This proof validation activity is essential to proof instruction because it can help students to develop the skills to assess their own learning while

constructing proofs (Selden & Selden, 2015a). However, students seldom encounter proof validation activities in mathematics classes as they are mainly exposed to correct proofs during instruction (Zerr & Zerr, 2011).

Peer-feedback is increasingly used in teacher-training courses, including mathematics education (e.g., Lavy & Shriki, 2014; Sluijsmans, Brand-Gruwel, Van Merriënboer, & Bastiaens, 2003), because it supports students' learning (Cho & Cho, 2011) and their assessment skills (Sluijsmans et al., 2003); the latter is a skill that every preservice teacher needs to develop. Preservice mathematics teachers particularly need to be able to assess proofs because most school mathematics curricula typically include them (Selden & Selden, 2015a). However, peer-feedback provision on proofs is likely to be challenging for preservice mathematics teachers. Studies on proofs showed that when undergraduate students are asked to validate proofs of different levels of correctness they could not reliably differentiate between correct and erroneous proofs (e.g., Inglis & Alcock, 2012; Selden & Selden, 2003), and that erroneous proofs are more challenging to the students to validate (Inglis & Alcock, 2012; Zerr & Zerr, 2011). Nevertheless, students' accuracy in proof validation seems to depend on the type of error in the proof (Sommerhoff, Ufer, & Kollar, 2016).

These findings are in line with peer-feedback research revealing that

* Corresponding author. LMU Munich, Department of Psychology, Leopoldstr. 44, 80802, Munich, Germany.

E-mail addresses: m.j.n.alqassab@gmail.com (M. Alqassab), j.w.strijbos@rug.nl (J.-W. Strijbos), ufer@math.lmu.de (S. Ufer).

¹ Permanent address: House 1944, Road 644, Sanabis 406, Kingdom of Bahrain.

peer-feedback provided by a student is shaped by the quality of the peer solution (e.g., Patchan & Schunn, 2015) which can be reflected in the type of error in the proof. Peer-feedback studies have been focusing on improving the content of peer-feedback provided by students through instructional scaffolds (e.g., Alqassab, Strijbos, & Ufer, 2018; Gielen, Peeters, Dochy, Onghena, & Struyven, 2010). Yet, this approach did not sufficiently work for peer-feedback providers with low domain knowledge (Alqassab et al., 2018). Such research needs to be informed by empirical studies investigating how peer-feedback providers deal with the peer solution during the peer-feedback provision process, thereby producing (in)accurate peer-feedback or learning outcomes (e.g., comprehension of the proof). Specifically, we need to simultaneously investigate the process of composing the peer-feedback message and its outcomes to better understand this complex activity. Accordingly, there is a need to explore how the quality of peer solution influences process measures such as cognitive processing of the peer solution during peer-feedback provision as well as outcome measures (peer-feedback content and proof comprehension) in order to deliver more efficient instructional support for preservice mathematics teachers during this activity.

Eye-tracking is a useful tool to infer cognitive processes during assessment-related activities based on the assumption that what is being attended to is also cognitively processed (the eye-mind assumption; Just & Carpenter, 1976). Previous eye-tracking studies that investigated proof validation (e.g., Inglis & Alcock, 2012) or processing of peer-feedback by recipients (e.g., Bolzer, Strijbos, & Fischer, 2015) provided insights into the elements of proofs heeded during proof validation and how cognitive processing of received peer-feedback is related to revision. However, no study—to our knowledge—has investigated cognitive processing during peer-feedback provision on peer solutions to proofs despite the need for process measures underlying the outcomes (i.e., peer-feedback content and proof comprehension) of this challenging activity.

A specific type of proofs that is often used as initial context in proof instruction is geometry proofs. Their usefulness is attributed to the 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) and making inferences from the figure in geometry proofs is assumed to be easier than making inferences from statements (Larkin & Simon, 1987). However, geometry proofs are still widely ignored in research on proof in mathematics education despite (a) the usefulness of figures for learning as exemplified in multimedia learning research in different domains with the help of eye-tracking methodology (for a review see Eitel & Scheiter, 2015) and (b) the emerging interest in implementing peer-feedback activities with preservice mathematics teachers on geometry proofs (e.g., Lavy & Shriki, 2014). Yet, empirical studies investigating how preservice mathematics teachers utilize the figure of the geometry proof during peer-feedback provision are still limited.

1.1. Geometry proof based on mental models

Dealing with geometry proofs requires multiple skills including deductive reasoning (Schoenfeld, 1986). Several psychological theories about deductive reasoning can be applied to proof construction (for reviews see Bara, Bucciarelli, & Lombardo, 2001; Stylianides & Stylianides, 2008). We use the Mental Model Theory (Johnson-Laird, Byrne, & Schaeken, 1992) because it has previously been extended to research on geometry proofs (Ufer, Heinze, & Reiss, 2009), and is frequently utilized in research on learning with text and figure (see Schnotz, 2002).

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 et al., 1992). The Mental Model Theory 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).

Ufer et al. (2009) extended the Mental Model Theory to explain reasoning processes underlying geometric proof construction. Their framework acknowledges that geometry proof tasks are often accompanied by a geometric figure. Mental models in this framework are not restricted to a specific geometric figure, but also comprises conceptual properties that define the geometric configuration (i.e., figural concept; see Fischbein, 1993). 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., problem text), and (b) visual information (i.e., figure). Students then generate intermediate conclusions in the second phase based on their mental model, which are then validated in the third phase by either trying to exclude contradicting alternative mental models, or by referring to a theorem that excludes the existence of such alternative models. Reading a geometry proof (attempt), thus, can be described by two different approaches: a text-based approach, that focuses on the different statements in the text and their mutual relations, or a figure-based approach, that focuses on what the statements given in the text mean in terms of the geometric configuration. The next section elaborates on how these two approaches can be employed during peer-feedback provision on geometry proofs.

1.2. Employing geometry proof mental models during peer-feedback provision

In the context of proofs, providing peer-feedback entails reading a proof attempt by a peer, reflecting on it, and judging its correctness (i.e., validation; Selden & Selden, 2015b). This act requires an involvement in all three phases of deductive reasoning described by the Mental Model Theory. In particular, the peer-feedback provider needs to construct a mental model (using information from the text and the figure) based on which s/he judges the correctness of the peer solution to produce peer-feedback.

Evidence from multimedia studies shows that a figure can support learning from text. For example, Eitel, Scheiter, Schüler, Nyström, and Holmqvist (2013) revealed that a figure acts as a mental scaffold that facilitates text-comprehension. Another study showed that the presence of pictures in items of a science test stimulated more efficient item-reading and better performance (Lindner, Eitel, Strobel, & Köller, 2017). Accordingly, we propose that adopting a figure-based mental model while providing peer-feedback on geometry proofs can facilitate feedback provision. Nevertheless, it is unclear under which conditions a figure-based approach is likely to be adopted. Unlike proof construction, in peer-feedback provision on geometry proofs the peer-feedback provider is presented with an already-constructed proof by a peer, thus the construction of the mental model is likely to be influenced by the quality of the written peer solution.

1.2.1. The role of peer solution quality in constructing mental models

During peer-feedback provision, the text of the peer solution represents the main body of the geometry proof. Hence, it is likely that the peer-feedback provider focuses on the text and inspects the figure in relation to the text. However, the quality of the peer solution might influence whether the peer-feedback provider adopts a figure-based or a text-based approach and to what extent s/he integrates both components of the peer solution to the geometry proof.

Eye-tracking studies suggest that when learning with text and figure (presented simultaneously), students focus mainly on the text and the processing of the figure is guided by the information available in the text (e.g., Eitel et al., 2013; Hegarty & Just, 1993; Stalbovs, Scheiter, &

Gerjets, 2015). However, there are also cases in which the figure can direct the text processing when presented earlier (for a review see Eitel & Scheiter, 2015). Nevertheless, for an activity like peer-feedback provision, the reader's focus is likely to be initially on the main component of the peer's proof (i.e., the text), which then would guide figure processing.

Ambiguous text, such as incomplete or inconclusive proof attempts, can lead to more utilization of the figure (Eitel et al., 2013). Yet, in cases that the text directs figure processing, the figure supports learning when enough information initially extracted from the text directs the extraction of further relevant information from the figure (Eitel & Scheiter, 2015; Hegarty & Just, 1993). Also, earlier context in the text influences how long readers look at later parts of the text (Rayner, 1998), with inconsistent information resulting in longer fixations (Hegarty, Mayer, & Green, 1992). Consequently, it can be expected that errors in the early parts of a peer solution to a geometry proof might lead to longer processing of the text and the figure, but due to the lack of information in the text, that guides extracting relevant information from the figure, the peer-feedback provider likely fails to adopt a figure-based approach which in turn might stimulate longer text processing of the later parts of a peer solution to a geometry proof.

1.3. Proof comprehension as a result of peer-feedback provision

Peer-feedback provision activities may lead to different outcomes. While the quality of the peer-feedback is more relevant for the learning of the recipients, the peer-feedback providers have the chance of deepening their knowledge about the learning task. When involved in peer-feedback provision on geometry proofs, proof comprehension is an important learning outcome to be measured for preservice mathematics teachers. Validating a peer's proof indeed involves reading and sense-making (i.e., comprehension) of the proof attempt (Selden & Selden, 2015a). Despite the eminent evidence regarding students' inability to validate proofs (e.g., Inglis & Alcock, 2012; Reiss, Heinze, & Klieme, 2000) research investigating the impact of the quality of the proofs being validated on proof comprehension is scarce.

Preservice mathematics teachers might be expected to learn from the mistakes of their peer(s) in the same manner as learning from erroneous worked-examples. However, given the challenges learners face with proofs and the mixed findings from erroneous worked-examples research (e.g., Große & Renkl, 2007; Isotani, Adams, Mayer, Durkin, Rittle-Johnson, & McLearen, 2011; Tsovaltzi et al., 2010), errors in the peer solution are likely to hinder proof comprehension. The peer-feedback providers' domain knowledge might play a role as learners with high domain knowledge seem to benefit more from erroneous worked-examples (Große & Renkl, 2007). Yet, there is also evidence that—regardless of their domain knowledge—learners did not benefit more from studying erroneous worked-examples compared to correct worked-examples (Isotani et al., 2011). It is therefore unclear whether errors in the peer solution will still hinder proof comprehension during peer-feedback provision after controlling for domain knowledge.

1.4. Peer-feedback content as a result of peer-feedback provision

Another aspect that cannot be ignored when investigating peer-feedback provision as a learning activity is the peer-feedback content because it is the main product of this activity. The peer-feedback content is usually defined in terms of its *type* (e.g., cognitive-verification, cognitive-elaboration, self-efficacy; Strijbos, Van Goozen, & Prins, 2012) or *accuracy* (i.e., correct/incorrect statements; Van Steendam, Rijlaarsdam, Sercu, & Van den Bergh, 2010). Unlike experts, peers may have inaccuracies in their peer-feedback that is not represented by the peer-feedback type. Peer-feedback content as a learning outcome for preservice mathematics teachers (Sluijsmans et al., 2003) is particularly important because they need to provide feedback to their future students that is of a useful type but also accurate. Accordingly, peer-

feedback content should be measured in terms of its type and accuracy.

Previous studies in other domains showed that the quality of the peer solution influenced peer-feedback types with low-quality written reports stimulating more elaborated peer-feedback, and high-quality reports stimulating more confirmatory peer-feedback (Cho & Cho, 2011; Patchan, Charney, & Schunn, 2009). It is unclear, however, how the quality of the peer solution affects the peer-feedback content (type & accuracy) when complex tasks such as geometry proofs are used.

1.5. Current study

The aim of this study was to investigate the impact of peer solution quality to a geometry proof task on: (a) peer-feedback providers' cognitive processing during peer-feedback provision, (b) peer-feedback providers' comprehension of the geometry proof, and (c) peer-feedback content. Accordingly, we analyzed preservice mathematics teachers' eye-movements (on text, figure, and transitions between text and figure) as indicators of cognitive processing (based on the eye-mind assumption; Just & Carpenter, 1976), and their proof comprehension and peer-feedback content as outcome measures of the peer-feedback provision activity. Two peer solutions of different quality (a near-correct and erroneous) were created. To test how errors in a peer solution influence cognitive processing of correct parts during peer-feedback provision, the peer solutions consisted of two parts of which the first was non-identical and the second identical. Specifically, the first part was correct for the near-correct peer solution (NCS), but had missing warrants for the erroneous peer solution (ERS). The second part was identical and contained very minor notation errors. We investigated the following hypotheses:

H1. Processing of peer solution

- *Text*: Compared to no errors, errors in the first part of the peer solution would lead to longer processing of the text of the first part of the peer solution (H1a), and to longer processing of the text of the second part of the peer solution (H1b), due to ambiguity of information (Hegarty et al., 1992) and the influence of earlier context (Rayner, 1998).
- *Figure*: Errors in the first part of the peer solution would lead to longer processing of the figure while assessing the first part of the peer solution compared to no errors (H1c) to resolve ambiguity in the text (Eitel et al., 2013), but to shorter processing of the figure while assessing the second part of the peer solution due to failure to construct a figure-based mental model (H1d) due to lack of information extracted earlier from the text (Hegarty & Just, 1993).
- *Transitions*: Compared to the NCS, more transitions from the text to the figure would occur for the ERS on the first part of the peer solution (H1e) to resolve ambiguity in the text (Eitel et al., 2013) and on the second part of the peer solution due to problems constructing a figure-based mental model (H1f) as an effect of the earlier text context (Rayner, 1998).

H2. Proof comprehension: Providing peer-feedback on the ERS would result in less comprehension of the geometry proof based on the findings that students have difficulties validating erroneous proofs (Reiss et al., 2000; Zerr & Zerr, 2011) and that they do not benefit from studying erroneous worked-examples compared to correct examples (Isotani et al., 2011).

H3. Peer-feedback content (type and accuracy): Based on the findings by Cho and Cho (2011) and Patchan et al. (2009), peer-feedback provided on the ERS would contain more cognitive-elaboration peer-feedback and self-efficacy peer-feedback, and peer-feedback provided on the NCS would contain more cognitive-surface peer-feedback and cognitive-verification peer-feedback (H3a). The peer-feedback provided on the ERS would be less accurate than that provided on the NCS because it is easier to identify correct geometry proofs (Reiss et al., 2000) (H3b).

2. Method

2.1. Participants and study design

Participants were fifty-three preservice mathematics teachers for high school (74% females; $M_{\text{age}} = 24.40$, $SD = 5.33$) from a German university. Participation in the study was voluntary and participants were compensated with €15 for their participation. All participants signed an informed consent before taking part. Participants were randomly assigned to one of two experimental conditions: NCS ($n = 27$) or ERS ($n = 26$). The experiment consisted of two phases. In phase 1, we used eye-tracking methodology to measure participants' cognitive processing of a peer solution while providing verbal peer-feedback on that solution. In phase 2, we measured participants' comprehension of the proof they provided peer-feedback on, their basic geometric knowledge, and background information such as age and gender. Initially, one additional participant was assigned to the ERS condition but had to be excluded due to inaccuracy of the eye-tracking data.

2.2. Materials and measures

2.2.1. Experimental task: peer solutions to a geometry proof

We used two fictional peer solutions to a geometry proof designed by a mathematics expert as having a non-identical first part and an identical second part (see Fig. 1). The peer solution consisted of nine steps in the NCS condition and of eight steps in the ERS condition. The non-identical part of the peer solutions consisted of steps 1–3 in the NCS and 1–2 in the ERS. The identical part consisted of the remaining steps (see Fig. 1). The design of the peer solutions was informed by prior research on the type of errors exhibited by students when performing geometry proofs (e.g., Reiss et al., 2000; Schoenfeld, 1986; Tapan & Arslan, 2009).

2.2.2. Eye-tracking measures

Total dwell time (in seconds) was used as a measure of cognitive processing during peer-feedback provision. Transitions from the text to the figural components of the peer solution were also used as a measure of cognitive processing due to the nature of the experimental task (i.e., geometry proof). The transitions indicate the degree to which the learner integrates information from the text and the figure (Mason, Tornatora, & Pluchino, 2013) and has been used in previous research (e.g., Bolzer et al., 2015; Mason et al., 2013; Stalbvos et al., 2015).

2.2.3. Peer-feedback instruction and example task

To control for participants' prior experience with peer-feedback, they were introduced to peer-feedback and its important role in supporting learning and they received information regarding constructive peer-feedback (i.e., 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 introduction, the participants received an example task consisting of a fictional peer solution to a geometry proof—similar to the experimental task—showing examples of constructive and non-constructive peer-feedback on that solution. For both the instruction and the experimental tasks, the participants were told that the solution had been generated by a peer.

2.2.4. Proof comprehension test

Participants' comprehension of the geometry proof was measured based on a model of reading comprehension of proofs (see Yang & Lin, 2008). Sixteen open-ended questions were created by an expert in mathematics (e.g., “In step 2 of the proof: Give the correct formulation of the theorem on triangles used in this step.”). One or two points were awarded for each correct assertion depending on the complexity of the question, and no points were awarded for missed or wrong assertions.

Two student assistants scored 10% of the tests reaching a good inter-rater agreement (Krippendorff's $\alpha = .80$). Subsequently, one student

assistant scored the remaining tests. The internal consistency was determined using Cronbach's α . Two questions had corrected item-total correlations $< .30$ and when removed Cronbach's α improved from .77 to .82. Parallel analysis followed by an Exploratory Factor Analysis supported a 1-factor solution, $\chi^2(104) = 126.15$, $p = .069$. Therefore, a total sum score was computed for the remaining items.

2.2.5. Basic geometric knowledge test

We measured students' basic geometric knowledge to control for between-group differences. The test consisted of 49 true/false items which were scored dichotomously as 0 if answered incorrectly and 1 if answered correctly (Cronbach's $\alpha = .79$) and measured different basic geometric knowledge, including properties of triangles, properties of a parallelogram, transversals, and quadrangles (e.g., “For every parallelogram, the opposite sides are parallel: true/false”). A total sum score of the 49 items was computed as a measure of participants' basic geometric knowledge.

2.3. Apparatus

The peer solutions were presented on a 22-inch screen with 1680×1050 -pixel resolution. Participants' eye-movements were recorded using a monocular head-mounted Dikablis 25 Hz eye-tracker (by Ergoneers LLC, Germany) with a 3-mm lens. The eye-tracker was calibrated for maximum accuracy in front of the screen before the start of phase 1.

2.4. Peer-feedback content: type and accuracy

Peer-feedback type was operationalized as having two dimension—purpose and style—based on a coding scheme by Strijbos et al. (2012). The purpose of peer-feedback could have (a) a cognitive focus on the content knowledge about 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, and (d) a self-efficacy focus on the providers' confidence in their ability to provide peer-feedback or to construct the proof. Furthermore, cognitive and metacognitive peer-feedback can each have two styles: verification or elaboration. Verification represents confirming or disconfirming statements about the correctness of the peer solution, or the learning approaches used to deal with the task. Elaboration represents detailed comments about parts of the peer solution that follow verifications or other elaborations in the form of correction, confirmation, justification, questioning, or suggestions. Cognitive peer-feedback can also contain cognitive surface comments about writing style (e.g., grammar and spelling mistakes).

Peer-feedback accuracy was computed by identifying the number of detected errors or correct statements by the peer-feedback provider. The same student assistants who coded for peer-feedback type also coded the accuracy of peer-feedback. The NCS had an extra type, and the ERS had more errors. Hence, the maximum peer-feedback accuracy score for the NCS was 14 points, whereas it was 16 points for the ERS. To compare the two conditions, proportional peer-feedback accuracy scores were computed.

2.5. Procedure

The participants were tested individually in a quiet room. Once the eye-tracker was calibrated with high accuracy, the participants received the peer-feedback instruction. Afterwards, they were presented with one of the fictional peer solutions and were instructed to provide verbal peer-feedback on that peer solution. They could navigate freely through different slides in phase 1. The participants' voices were recorded when they provided peer-feedback using an external voice recorder. The eye-tracker was removed at the end of phase 1. Phase 2 consisted of

Fig. 1. Experimental task (first row); fictional peer solutions for the NCS condition (second row) and the ERS condition (third row).

answering the proof comprehension test, the basic geometric knowledge test, and some background information. In phase 2, the participants received a printed copy of the peer solution they had provided peer-feedback on in phase 1 because the proof comprehension test asked specific questions related to each step of the proof and did not ask for reproducing parts of the proof. They completed both phases at their own pace and the entire experiment lasted between 90 and 120 min.

2.6. Analyses

2.6.1. Eye-tracking data pre-processing

Total dwell time and transitions were computed for predefined Areas of Interest (AOI)s for each condition. The figure was the same for

both conditions and did not change in the first (non-identical) and second (identical) part of each solution. However, since the participants looked at the figure in relation to the 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) the text of the first (non-identical) peer solution part, (2) the figure corresponding to the first peer solution part, (3) the text of the second (identical) peer solution part, and (4) the figure corresponding to the second peer solution part. Transitions from the text to the figure were computed for each part of the peer solution.

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; a procedure used in previous studies (e.g., Bednarik & Tukiainen, 2008; Bolzer et al., 2015). The transitions were also divided by the total dwell time as the two measures were correlated for the first part ($r = .83$, $CI = [0.90, 0.74]$, $p < .001$) and the second part of the peer solutions ($r = .71$, $CI = [0.82, 0.55]$, $p < .001$).

Since parts (first and second) and components (text and figure) of the peer solutions had different AOI sizes, we divided the PTDT and the transitions by the size of AOIs in pixels. This made it possible to compare the eye-tracking measures on different parts of the peer solution between-subjects and within-subjects.

2.6.2. Peer-feedback coding

The voice-recorded peer-feedback was transcribed and segmented following Strijbos, Martens, Prins, and Jochems's (2006) approach, with the smallest meaningful segment as the unit of analysis. Two student assistants independently segmented 10% of the data each time over two rounds until an acceptable percentage agreement was reached (92.5% lower bound; 93% upper bound). Afterwards, one of the student assistants segmented the remaining data. Two student assistants independently coded 10% of the data each time over five rounds until an acceptable inter-rater reliability was reached for peer-feedback type (Krippendorff's $\alpha = .81$) and accuracy (Krippendorff's $\alpha = .76$). The same student assistants then coded half of the data each. No metacognitive peer-feedback was identified in the current sample, and participants in the ERS condition did not provide affective peer-feedback.

2.6.3. Statistical tests

Robust statistical methods based on Wilcox (2012) were used to analyze the eye-tracking data due to violation of normality and homogeneity of variance assumptions. Between-within-subjects ANOVAs on trimmed means were used (for details see Mair & Wilcox, 2015; Wilcox, 2012). Yuen's bootstrap version t-tests on trimmed means were used for post hoc analyses as proposed by Wilcox (2012). A robust heteroscedastic measure of effect size is the *explanatory measure of effect size* (ξ) proposed by Wilcox and Tian (2011) that represents a strength of association. Values of .15, .35, and .50 correspond to small, medium, and large respectively (Wilcox & Tian, 2011). For the main and interaction effects, we report the test statistics (Q) and p -values. For post hoc tests we report the test statistics, 95% confidence intervals (CI) of the trimmed mean difference, and ξ .

To our knowledge, no non-parametric test allows the inclusion of a covariate with mixed-analyses. Therefore, we ran M/ANCOVAs only for research questions concerning between-subject factors (i.e., proof comprehension and peer-feedback content). Wilcox robust ANCOVA was used whenever an assumption of parametric tests was violated. Four design points were used that were automatically specified by the function with a minimum of 12 participants in each group around the design points (Wilcox, 2012). We report the adjusted p -values with Bonferroni correction. For the robust Wilcox ANCOVAs, we report Q_x for each design point (where x is the design point) and the 95% CI . For the M/ANCOVAs, partial eta squared (η_p^2) is reported as a measure of effect size.

2.6.4. Software

The R package for Wilcox Robust Estimation and Testing (WRS2 v. 0.3–2; Mair & Wilcox, 2015) was used to run the robust analyses. We used the R Project for Statistical Computing (version 3.3.1) for the robust analyses, and IBM SPSS Statistics 23 to run the remainder of the analyses.

3. Results

3.1. Data screening

The standardized skewness and kurtosis values were determined for

each research condition separately and were within the acceptable range (± 3 ; Tabachnick & Fidell, 2013) for proof comprehension and peer-feedback accuracy. The eye-tracking measures (i.e., PTDT and transitions) had similarly acceptable standardized skewness and kurtosis values, apart from the PTDT on text ($z_{\text{skewness}} = 3.20$) and on figure ($z_{\text{skewness}} = -3.20$) for the first part of the peer solution in the NCS condition. The standardized skewness and kurtosis values were outside the acceptable range for the basic geometric knowledge ($z_{\text{skewness}} = -4.00$, $z_{\text{kurtosis}} = 4.50$) in the NCS condition, and for peer-feedback types including cognitive-surface peer-feedback (NCS: $z_{\text{skewness}} = 3.60$; ERS: $z_{\text{skewness}} = 5.02$; $z_{\text{kurtosis}} = 6.61$), and self-efficacy peer-feedback (NCS: $z_{\text{skewness}} = 7.36$; $z_{\text{kurtosis}} = 14.63$; ERS: $z_{\text{skewness}} = 4.48$). Two extreme univariate outliers ($z > |3.29|$; Field, 2009) were identified; one for basic geometric knowledge and one for cognitive-surface peer-feedback. 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 value of the other outlier did not improve the standardized skewness and kurtosis, so we retained the original values. No multivariate outliers were identified.

3.2. Manipulation check

To ensure that participants in each condition did not differ regarding their basic geometric knowledge, we performed an independent samples t -test. There was no significant difference in basic geometric knowledge between the NCS condition ($M = 42.26$, $SD = 3.99$) and the ERS condition ($M = 41.46$, $SD = 3.71$), $t(51) = 0.75$, $p = .455$, $d = 0.13$.

3.3. Eye-movements

3.3.1. Proportional total dwell time on peer solution's text

Mixed ANOVA with peer solution quality as a between-subjects factor, peer solution part (first vs. second) as a within-subjects factor, and PTDT on the text as a dependent variable, revealed a significant main effect for peer solution quality, $Q = 129.35$, $p < .001$, no significant main effect for part of peer solution, $Q = 0.22$, $p = .645$, and a significant interaction effect, $Q = 31.54$, $p < .001$. Post hoc comparisons with Bonferroni correction revealed that participants in the NCS condition had shorter PTDT on the text of the first part of the peer solution than the second part, $t(16) = -8.35$; $CI = [-.0002, -.0001]$; $p < .001$; $\xi = .88$. Conversely, participants in the ERS condition had longer PTDT on the text of the first part of the peer solution than on the second part, $t(15) = 3.20$; $CI = [.0001, .0004]$; $p = .024$; $\xi = .61$, explaining the significant interaction (see Fig. 2). However, participants in the ERS condition had significantly longer PTDT on the text of the first part, $t(15.66) = 9.54$; $CI = [-.0009, -.0006]$; $p < .001$; $\xi = .95$, as well as on the text of the second part, $t(30.86) = 10.00$; $CI = [-.0004, -.0003]$; $p < .001$; $\xi = .92$, than participants in the NCS condition (Fig. 2).

3.3.2. Proportional total dwell time on peer solution's figure

Mixed ANOVA with peer solution quality as a between-subjects factor, peer solution part (first vs. second) as a within-subjects factor, and PTDT on the figure as a dependent variable revealed a significant main effect of quality of peer solution, $Q = 32.27$; $p < .001$, a significant main effect of peer solution part, $Q = 65.08$; $p < .001$, and a significant interaction, $Q = 578.15$; $p < .001$. Post hoc comparisons with Bonferroni correction revealed that participants in the NCS condition had shorter PTDT on the figure while assessing the first part of the peer solution than those in the ERS condition, $t(16.81) = 20.92$; $CI = [-.0014, -.0011]$; $p < .001$; $\xi = .96$. However, for the second peer solution part, participants in the NCS condition had longer PTDT

Fig. 2. Proportional total dwell time on the text of the first and the second peer solution parts for the NCS condition and the ERS condition. Error bars are standard errors of means.

Fig. 3. Proportional total dwell time on the figure of the first and the second peer solution parts for the NCS condition and the ERS condition. Error bars are standard errors of means.

on the figure than participants in the ERS condition, $t(30.86) = 10.02$; $CI = [.0006, .0008]$; $p < .001$; $\xi = .92$ (see Fig. 3).

3.3.3. Transitions from the text to the figure

Mixed ANOVA with peer solution quality as the between-subjects factor, peer solution part as the within-subjects factor, and transitions from the text to the figure as the dependent variable revealed a significant main effect of peer solution quality, $Q = 27.46$; $p < .001$, a significant main effect of part of peer solution, $Q = 36.42$; $p = .000$, and a significant interaction effect, $Q = 93.51$; $p < .001$.

Post hoc comparisons with Bonferroni correction showed that participants in the ERS condition had significantly more transitions on the first peer solution part, $t(18.05) = 8.51$; $CI = [-.0002, -.0001]$; $p < .001$; $\xi = .98$, but there was no significant difference between both conditions in the transitions on the second peer solution part, $t(25.26) = 1.37$; $CI = [.000, .0001]$; $p = .730$; $\xi = .28$ (see Fig. 4). Participants in the erroneous peer solution condition had significantly more transitions on the first than on the second part of peer solution, $t(16) = -3.09$; $CI = [-.0001, 0]$; $p = .028$; $\xi = .57$. Conversely, participants in the NCS condition had significantly less transitions on the first than on the second part of peer solution, $t(15) = 9.71$; $CI = [.0001, .0002]$; $p < .001$; $\xi = .94$, explaining the interaction (see Fig. 4).

3.4. Differences in proof comprehension

A one-way ANCOVA with quality of peer solution as an independent variable, proof comprehension scores as a dependent variable, and

Fig. 4. Transitions from the text to the figure of the first and the second peer solution parts for the NCS condition and the ERS condition. Error bars are standard errors of means.

Fig. 5. Proof comprehension scores for the NCS condition and the ERS condition. Error bars are standard errors of means.

basic geometric knowledge as a covariate, revealed a significant effect of peer solution quality on proof comprehension, after controlling for basic geometric knowledge, $F(1, 50) = 4.99$, $p = .030$, $\eta_p^2 = .09$. As illustrated in Fig. 5, participants who provided peer-feedback on the NCS performed significantly better in the proof comprehension test than those who provided peer-feedback on the ERS. The covariate basic geometric knowledge was significantly related to providers' proof comprehension, $F(1, 50) = 30.44$, $p < .001$, $\eta_p^2 = .38$.

3.5. Type of peer-feedback content

To investigate the impact of the quality of the peer solution on different peer-feedback types (cognitive-surface, cognitive-verification, cognitive-elaboration, and self-efficacy), we ran a one-way MANCOVA. The assumption of homogeneity of variance was met for cognitive-surface peer-feedback ($p = .785$), cognitive-elaboration peer-feedback ($p = .094$), and self-efficacy peer-feedback ($p = .087$), but it was violated for cognitive-verification peer-feedback ($p = .008$). Thus, cognitive-verification peer-feedback was not included in the MANCOVA. The assumptions of the homogeneity of covariance matrices, $p = .188$, and homogeneity of regression slopes, $p = .900$, were met.

Using Pillai's trace, there was no significant effect of quality of peer solution on cognitive-surface peer-feedback, cognitive-elaboration peer-feedback, and self-efficacy peer-feedback after controlling for the covariate basic geometric knowledge, $V = .01$, $F(3, 48) = .13$, $p = .940$, $\eta_p^2 = .01$ (see Fig. 6). The covariate basic geometric knowledge was not related to peer-feedback types, $V = .04$, $F(3, 48) = .68$, $p = .569$, $\eta_p^2 = .04$.

Wilcox robust ANCOVA was conducted for cognitive-verification peer-feedback. The findings revealed no significant impact of peer solution quality on the amount of provided cognitive-verification peer-

Fig. 6. Proportions of peer-feedback types for the NCS condition and the ERS condition: CS = cognitive-surface, CV = cognitive-verification, CE = cognitive-elaboration, and SE = self-efficacy. Error bars are standard errors of means.

feedback after controlling for basic geometric knowledge, $Q_{42} = 0.07$, $CI = [-.14, .13]$, $p = .947$; $Q_{43} = 0.24$, $CI = [-.12, .10]$, $p = .816$; $Q_{44} = 0.12$, $CI = [-.15, .14]$, $p = .905$; $Q_{45} = 0.47$, $CI = [-.20, .15]$, $p = .646$ (see Fig. 6).

3.6. Accuracy of peer-feedback content

A one-way ANCOVA with quality of peer solution as an independent variable, peer-feedback accuracy as a dependent variable, and basic geometric knowledge as a covariate, revealed a significant impact of peer solution quality on peer-feedback accuracy, after controlling for basic geometric knowledge, $F(1, 50) = 7.69$, $p = .008$, $\eta_p^2 = .13$. Participants in the NCS condition provided significantly more accurate peer-feedback than participants in the ERS condition (see Fig. 7). The covariate basic geometric knowledge also had a significant impact on peer-feedback accuracy, $F(1, 50) = 10.31$, $p = .002$, $\eta_p^2 = .17$.

3.7. Links between eye-movements and outcome measures

As a post hoc analysis, we investigated whether the adoption of a figure-based mental model resulted in better outcomes (proof comprehension, peer-feedback accuracy) through linear regressions. Due to the

Fig. 7. Proportions of peer-feedback accuracy for the NCS condition and the ERS condition. Error bars are standard errors of means.

small sample size and the increased risk of type I error inflation as a result of multiple testing and complex models, this was done only for the condition with better proof comprehension and peer-feedback accuracy (i.e., NCS condition) to investigate whether more efficient (figure-based) processing in the NCS condition predicts proof comprehension and peer-feedback accuracy after controlling for prior geometric knowledge. Also, we used Bonferroni correction to control the error rate resulting from additional tests with proof comprehension and peer-feedback accuracy. Two multiple linear regressions were calculated to predict proof comprehension and peer-feedback accuracy based on PTDT on the figure of the second part of peer solution controlling for basic geometric knowledge. The regression model was significant for proof comprehension, $F(2, 24) = 7.83$, $p = .010$, $R^2 = .34$, and peer-feedback accuracy, $F(2, 24) = 7.77$, $p = .010$, $R^2 = 0.34$. PTDT on the figure was a significant predictor of peer-feedback accuracy ($\beta = 0.43$, $p = .049$) but not of proof comprehension ($\beta = .11$, $p = 1$).

4. Discussion and conclusion

This study aimed to identify the impact of peer solution quality on peer-feedback provision. In particular, we investigated how the quality of a peer solution to a geometry proof—near correct solution (NCS) vs. erroneous solution (ERS)—influenced the peer-feedback providers' cognitive processing during peer-feedback provision, their proof comprehension, and the content of their peer-feedback. Based on previous research on (geometry) proofs (e.g., Inglis & Alcock, 2012; Reiss et al., 2000; Sommerhoff et al., 2016), multimedia learning (e.g., Eitel et al., 2013; Hegarty & Just, 1993; Stalbvos et al., 2015), and peer-feedback (e.g., Cho & Cho, 2011; Patchan et al., 2009), we hypothesized that errors in the first part of the peer solution would lead to longer PTDT spent at the text and figure of the peer solution, but would prevent the adoption of a figure-based mental model during peer-feedback provision on the second part, and it would lead to more transitions from the text to the figure, and to less comprehension of the proof. Additionally, we expected that errors in the peer solution would result in more cognitive-elaboration and self-efficacy peer-feedback, whereas the absence of errors would result in more cognitive-verification and cognitive-surface peer-feedback. Finally, we expected that errors in the peer solution would lead to less accurate peer-feedback.

4.1. The impact of peer solution quality on the processing of the peer solution

Participants who provided peer-feedback on the ERS had significantly more time on the text of the first part (non-identical) and the second part (identical) of the peer solution compared to those who provided peer-feedback on the NCS (H1a & H1b were supported). Also, participants in the ERS condition had significantly longer PTDT on the figure while assessing the first part, but less PTDT on the figure while assessing the second part of the peer solution compared to participants in the NCS condition (H1c & H1d were supported). These findings are consistent with previous research showing that ambiguous arithmetic problems result in longer fixations (Hegarty et al., 1992) and that ambiguous text leads to more utilization of the figure (Eitel et al., 2013). Since a geometric figure contains more properties that are not specifically stated in the premises (Koedinger & Anderson, 1990), it seems that the participants needed to check the correctness of statements against the figure in the absence of warrants in the first part of the ERS. This is also supported by the finding that the participants in the ERS condition showed significantly more transitions from the text to the figure of the first part of the peer solution than those in the NCS condition (H1e was supported) in an attempt to integrate both types of information (Mason et al., 2013).

The absence of errors and the availability of warrants in the first part of the NCS seemed to facilitate the adoption of a figure-based approach to assess the second part of the peer solution. By adopting the

figure-based approach, the participants in the NCS condition did not gaze at the text as long as the participants in the ERS condition, and rather focused on the figure while providing peer-feedback on the second part of the peer solution. Conversely, encountering errors in the first part of the peer solution—more specifically missing warrants—stimulated a text-based approach. Therefore, participants in the ERS condition had significantly more PTDT on the text of the second peer solution part while providing peer-feedback than participants in the NCS condition. These findings are consistent with previous findings suggesting that the information extracted earlier from the text guides the extraction of relevant information from the figure (Hegarty & Just, 1993) and show that the same applies during peer-feedback provision on tasks with text and figure components. Although errors early on in the peer solution lead to longer PTDT on the first part than the second part, participants in the ERS condition still spent significantly longer PTDT on the second part compared to participants in the NCS condition, supporting the finding that earlier text context stimulates the processing of later text context (Rayner, 1998). Taken together, the findings of this study not only support that information in the text guides the utilization of the figure during peer-feedback provision, but also suggest that the ambiguity of earlier parts of the text can lead to even more utilization of the text, probably due to failure to extract relevant information from the figure (Eitel & Scheiter, 2015).

We expected that the failure to construct a figure-based mental model for the ERS condition would result in more transitions from the text to the figure on the second peer solution part compared to the NCS condition, but we did not find evidence for this hypothesis (H1f was rejected). Our findings showed an increase in transitions for the NCS condition on the second part, and a decrease for the ERS condition. This might be explained by the nature of the task, as the participants in the NCS still had to check the correctness of the text, and thus adopting a figure-based approach might have required more transitions from the text to the figure later on.

However, an alternative explanation might be that the participants in ERS condition had longer PTDT on the figure than the participants in the NCS condition when providing peer-feedback on the first peer solution part. Consequently, these participants did not need to transition from the text to the figure more frequently later on. If so, no significant differences would be observed between both conditions in the PTDT on the text of the second part as this was identical for both conditions. Yet, our results showed that the participants in the ERS condition still had significantly longer PTDT on the text of the second peer solution part than those in the NCS condition.

4.2. The impact of peer solution quality on proof comprehension

Participants in the NCS condition had a better comprehension of the geometry proof compared to participants in the ERS condition (H2 was supported). This could indicate that the adoption of a figure-based approach (for the NCS condition as found by the eye-tracking data) facilitates proof comprehension during peer-feedback provision on geometry proofs. However, we did not find evidence that it predicted proof comprehension in our study. Thus, this assumption requires further investigation in future studies.

The difference in proof comprehension between the two conditions suggests that errors in a peer solution may inhibit peer-feedback providers' comprehension of the proof. This is consistent with the findings by Isotani et al. (2011) who showed that middle-school students did not benefit from studying erroneous worked-examples when learning decimals. Although several studies reported that studying erroneous worked-examples resulted in cognitive gains for students (e.g., Große & Renkl, 2007; Tsovaltzi et al., 2010), the participants in our study did not seem to better comprehend the proof while providing peer-feedback on the ERS. One reason might be that geometry proofs are challenging for students to validate (Reiss et al., 2000). Furthermore, unlike our study, previous studies reporting a positive effect of erroneous worked-

examples often combined the worked-example with an additional instructional support such as self-explanation (e.g., Große & Renkl, 2007) or error detection help (e.g., Tsovaltzi et al., 2010). Hence, preservice mathematics teachers might only benefit from providing peer-feedback on erroneous geometry proofs with the help of additional instructional support or via repeated exposure to such activity, but this should be examined in future research.

4.3. The impact of the peer solution quality on peer feedback content

The type of peer-feedback that the participants provided (cognitive-surface, cognitive-verification, cognitive-elaboration, and self-efficacy) did not differ depending on the quality of the peer solution (H3a was rejected). This is consistent with Patchan and Schunn (2015) who found that the amount of different types of peer-feedback (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 peer-feedback, our study revealed that peer-feedback provided by preservice mathematics teachers on the NCS to a geometry proof was more accurate than that provided on the ERS (H3b was supported). This supports findings from studies with high school students (grades 7 and 13; Reiss et al., 2000) and undergraduate mathematics students (e.g., Inglis & Alcock, 2012) that showed that judging the correctness of erroneous proofs was more challenging for students compared to correct proofs. A similar finding was also reported by Zerr and Zerr (2011), as undergraduates in their study appeared to—descriptively—be more successful at assessing correct peer solutions than erroneous peer solutions. It seems that even preservice mathematics teachers have difficulties assessing erroneous geometry proofs compared to correct proofs, which might be due to their limited experience with this activity (Zerr & Zerr, 2011). Whether preservice mathematics teachers can become better at assessing erroneous proofs as they accumulate more experiences is an issue that requires further investigation.

Nevertheless, our study suggests that adopting a figure-based mental model during peer-feedback provision (on the NCS of geometry proof) can facilitate peer-feedback provision as PTDT spent looking at the figure predicted more accurate peer-feedback for the NCS condition. Accordingly, whether peer-feedback provision skills on geometry proofs (and proof validation) can be improved by supporting more efficient utilization of the figure component through external support such as implementation intentions (Stalbovs et al., 2015) is worth investigation in future studies.

4.4. Methodological limitations and future directions

Among the limitations of our study is that it relied on eye-tracking measures that do not provide information about failure in processing. Future research could add cued-retrospective reports to examine failure (s) in processing (Hyönä, 2010), and whether longer PTDT is a valid measure of cognitive processing while providing peer-feedback. The small sample size in our study also indicates that the findings should be treated as explorative and additional studies should attempt to replicate our findings. Additionally, we used fictional peer solutions to have a stronger control on the effect of the quality of the peer solution which might not have exactly represented the solution produced by a peer in the natural settings. Yet, fictional scenarios can invoke reactions comparable to real situations (Robinson & Clore, 2001), and the peer-solutions in our study were designed based on previous findings regarding how students perform these tasks (e.g., Reiss et al., 2000; Schoenfeld, 1986; Tapan & Arslan, 2009).

Basic geometric knowledge was controlled for when comparing the two conditions because it was previously found to predict performance on geometry proofs (Ufer et al., 2009). Nevertheless, there are other important factors that our study did not take into account such as metacognitive skills and the ability to construct geometry proofs.

Finally, we studied peer-feedback provision with preservice mathematics teachers using a very specific task (i.e., geometry proof) with specific errors (missing warrants) in the first part of the solution, thus our findings should be further investigated with different samples, type and position of errors, and other learning tasks.

4.5. Practical implications

Our study suggests that errors in a peer solution can influence mental models (figure vs. text) construction during peer-feedback provision on geometry proofs. Preservice mathematics teachers are more capable of comprehending a geometry proof and providing more accurate peer-feedback when they provide peer-feedback on a solution with few errors indicating the difficulty of such a task for preservice (mathematics) teachers. However, the adoption of a figure-based mental model seems to facilitate peer-feedback provision on these tasks. Therefore, preservice (mathematics) teachers need to receive external support such as error detection help (e.g., Tsovaltzi et al., 2010), peer-feedback training (e.g., Sluijsmans et al., 2003), or implementation intentions (Stalbovs et al., 2015) to help them deal with the complexity of the task and utilize different sources of information (text vs. figure) more efficiently.

Acknowledgments

The project was funded by the Elite Network of Bavaria, an initiative of the Bavarian Ministry of Education, Science and the Arts (Project number: K-GS-2012-209).

References

- Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, 33(1), 11–30. <https://doi.org/10.1007/s10212-017-0342-0>.
- Bara, B. G., Bucciarelli, M., & Lombardo, V. (2001). Model theory of deduction: A unified computational approach. *Cognitive Science*, 25(6), 839–901.
- 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. <https://doi.org/10.1145/1344471.1344497>.
- 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: Peer feedback processing with eye tracking. *Journal of Computer Assisted Learning*, 31(5), 422–434. <https://doi.org/10.1111/jcal.12091>.
- Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, 39, 629–643. <https://doi.org/10.1007/s11251-010-9146-1>.
- Eitel, A., & Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review*, 27(1), 153–180. <https://doi.org/10.1007/s10648-014-9264-4>.
- Eitel, A., Scheiter, K., Schüller, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction*, 28, 48–63. <https://doi.org/10.1016/j.learninstruc.2013.05.002>.
- 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.
- Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4), 304–315. <https://doi.org/10.1016/j.learninstruc.2009.08.007>.
- 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. <https://doi.org/10.1016/j.learninstruc.2007.09.008>.
- Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. Lester (Ed.). *Second handbook of research on mathematics teaching and learning* (pp. 805–842). Greenwich, CT: Information Age Publishing.
- Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, 32(6), 717–742. <https://doi.org/10.1006/jmla.1993.1036>.
- 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. <https://doi.org/10.1037/0022-0663.84.1.76>.
- Hodds, M., Alcock, R., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, 45(1), 62–101. <https://doi.org/10.5951/jresmetheduc.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.
- Hyönä, J. (2010). The use of eye-movements in the study of multimedia learning. *Learning and Instruction*, 20(2), 172–176. <https://doi.org/10.1016/j.learninstruc.2009.02.013>.
- Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43(4), 358–390. <https://doi.org/10.5951/jresmetheduc.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.). *Sixth European conference on technology enhanced Learning: Towards ubiquitous learning* (pp. 181–195). Palermo, Italy: Springer.
- Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, 99(3), 418–439.
- Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processing. *Cognitive Psychology*, 8(4), 441–480. [https://doi.org/10.1016/0010-0285\(76\)90015-3](https://doi.org/10.1016/0010-0285(76)90015-3).
- Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14(4), 511–550.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–100.
- Lavy, I., & Shriki, A. (March 2014). Engaging prospective teachers in peer assessment as both assessors and assesses: The case of geometrical proofs. *International Journal for Mathematics Teaching and Learning*, 1–32 Retrieved from: <http://www.cimt.plymouth.ac.uk/jour-nal/lavey2pdf>.
- Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and Instruction*, 47, 91–102. <https://doi.org/10.1016/j.learninstruc.2016.10.007>.
- Mair, P., & Wilcox, R. (2015). Robust statistical methods: The R package WRS2. Retrieved from <http://cran.usth.edu/web/packages/WRS2/vignettes/WRS2.pdf>.
- Mason, L., Tornatora, M. C., & Pluchino, P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers & Education*, 60(1), 95–109. <https://doi.org/10.1016/j.compedu.2012.07.011>.
- National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- 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. <https://doi.org/10.17239/jowr-2008.01.01.1>.
- 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. <https://doi.org/10.1007/s11251-015-9353-x>.
- Rayner, K. (1998). Eye-movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422. <https://doi.org/10.1037/0033-2909.124.3.372>.
- 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.
- Robinson, M. D., & Clore, G. L. (2001). Simulation, scenarios, and emotional appraisal: Testing the convergence of real and imagined reactions to emotional stimuli. *Personality and Social Psychology Bulletin*, 27(11), 1520–1532. <https://doi.org/10.1177/01461672012711012>.
- Schnotz, W. (2002). Commentary: Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, 14(1), 101–120. <https://doi.org/10.1023/A:1013136727916>.
- 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.
- 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. <https://doi.org/10.2307/30034698>.
- Selden, A., & Selden, J. (2015a). Validations of proofs as a type of reading and sense-making. In K. Beswick, T. Muir, & J. Wells (Vol. Eds.), *Proceedings of the 39th conference of the international group for the psychology of mathematics education: Vol. 4*, (pp. 145–152). Hobart, Tasmania: PME.
- Selden, A., & Selden, J. (2015b). 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*.
- Sluijsmans, D. M. A., Brand-Gruwel, S., Van Merriënboer, J. J. G., & Bastiaens, T. J. (2003). The training of assessment skills to promote the development of reflection skills in teacher education. *Studies In Educational Evaluation*, 29(1), 23–42. [https://doi.org/10.1016/S0191-491X\(03\)90003-4](https://doi.org/10.1016/S0191-491X(03)90003-4).
- Sommerhoff, D., Ufer, S., & Kollar, I. (2016). Proof validation aspects and cognitive student prerequisites in undergraduate mathematics. In C. Csikos, A. Rausch, & J. Sztányi (Vol. Eds.), *Proceedings of the 40th conference of the international group for the psychology of mathematics education: Vol. 4*, (pp. 219–226). Szeged, Hungary: PME.
- Stalbovs, K., Scheiter, K., & Gerjets, P. (2015). Implementation intentions during multimedia learning: Using if-then plans to facilitate cognitive processing. *Learning and Instruction*, 35, 1–15. <https://doi.org/10.1016/j.learninstruc.2014.09.002>.
- 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. <https://doi.org/10.1016/j.compedu.2005.04.002>.
- Strijbos, J. W., Van Goozen, B., & Prins, F. (2012, August). Developing a coding scheme for analysing peer feedback messages. *Paper presented at the EARL-SIG 1 assessment*

- and evaluation conference, Brussels, Belgium.
- Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, 38(3), 289–321. <https://doi.org/10.2307/30034869>.
- Stylianides, G. J., & Stylianides, A. J. (2008). Proof in school mathematics: Insights from psychological research into students' ability for deductive reasoning. *Mathematical Thinking and Learning*, 10(2), 103–133.
- 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 A*, 6(3), 54–60.
- Tsovaltzi, D., Melis, E., McLaren, B. M., Meyer, A. K., Dietrich, M., & Gogvadze, 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 (Vol. Eds.), *Sustaining TEL: From innovation to learning and practice: Vol. 3683*, (pp. 357–373). Berlin/Heidelberg: Springer.
- Ufer, S., Heinze, A., & Reiss, K. (2009). Mental models and the development of geometric proof competency. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Vol. 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 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. <https://doi.org/10.1016/j.learninstruc.2009.08.009>.
- Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Oxford, England: Elsevier.
- Wilcox, R. R., & Tian, T. S. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics*, 38(7), 1359–1368. <https://doi.org/10.1080/02664763.2010.498507>.
- Yang, K. L., & Lin, F. L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, 67(1), 59–76. <https://doi.org/10.1007/s10649-007-9080-6>.
- Zerr, J. M., & Zerr, R. J. (2011). Learning from their mistakes: Using student's incorrect proofs as a pedagogical tool. *Primus*, 21(6), 530–544. <https://doi.org/10.1080/105119709033869>.