Scaffolding peer-assessment skills

Citation for published version (APA):

Konings, K. D., van Zundert, M., & van Merriënboer, J. J. G. (2019). Scaffolding peer-assessment skills: Risk of interference with learning domain-specific skills? *Learning and Instruction*, 60, 85-94. <https://doi.org/10.1016/j.learninstruc.2018.11.007>

Document status and date:

Published: 01/04/2019

DOI:

[10.1016/j.learninstruc.2018.11.007](https://doi.org/10.1016/j.learninstruc.2018.11.007)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Scaffolding peer-assessment skills: Risk of interference with learning domain-specific skills?

Karen D. Könings*, Marjo van Zundert, Jeroen J.G. van Merriënboer

Maastricht University, Department of Educational Development & Research and Graduate School of Health Professions Education, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

ARTICLE INFO

Keywords:

Peer assessment
Task complexity
Scaffolding
Instruction
Cognitive load

ABSTRACT

Giving students complex learning tasks combined with peer-assessment tasks can impose a high cognitive load. Scaffolding has proven to reduce cognitive load during learning and improve accuracy on domain-specific tasks. This study investigated whether scaffolding has a similar, positive effect on the learning of peer-assessment tasks. We hypothesised that: (1) domain-specific scaffolding improves domain-specific accuracy and reduces time on task and perceived mental effort, and (2) peer-assessment scaffolding improves peer-assessment accuracy and reduces time on task and perceived mental effort. Additionally, we explored whether there was an interaction between domain-specific and peer-assessment scaffolding. In a 2x2 experiment with the factors domain-specific scaffolding (present, absent) and peer-assessment scaffolding (present, absent), 236 secondary school students assessed the performance of fictitious peers in an electronic learning environment. We found that domain-specific accuracy indeed improved with domain-specific scaffolding, confirming our first hypothesis. Our tests of the second hypothesis, however, revealed surprising results: peer-assessment scaffolding significantly increased accuracy and mental effort during learning, it had no effect on peer-assessment accuracy at the test and led to reduced domain-specific accuracy, even when combined with domain-specific scaffolding. These results suggest that scaffolding students' peer assessment before they have mastered the task at hand can have disturbing effects on students' ability to learn from the task.

1. Introduction

In peer assessment, students evaluate the performance of their fellow students (Topping, 2009). More specifically, peer assessment is an 'arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status' (Topping, 1998, p. 250). Although the growing popularity of peer assessment in education has already triggered a vast body of research (for an overview, see Ashenafi, 2017; Van Zundert, Sluijsmans, & Van Merriënboer, 2010), studies on peer-assessment skills typically choose the *assessee*, that is, the receiver of the peer feedback, as their object of study. As such, these studies examine how feedback from the assessor (the feedback giver) influences the *assessee's* use of feedback and learning (e.g., Cheng, Liang, & Tsai, 2015; Demiraslan Çevik, Haşlamani, & Çelik, 2015; Gielen, Tops, Dochy, Ongehe, & Smeets, 2010; Patchan, Hawk, Stevens, & Schunn, 2013; Strijbos, Narciss, & Dünnebier, 2010). As yet, knowledge for *assessors* of how to adequately teach students to perform peer assessments (Van Popta,

Kral, Camp, Martens, & Simons, 2017) is scant and the studies that are available focus essentially on providing tools for peer assessment rather than on students' learning of peer-assessment skills. Our study concentrates on the peer assessor and investigates whether scaffolding of peer-assessment skills has positive effects on their development in the same way that domain-specific scaffolding affects the development of domain-specific skills.

Earlier initiatives to support students in their task to assess the performance of a peer include prompting assessors to describe what the peer did or did not do correctly and give suggestions for improvement. This has shown to be effective in improving the quality of the feedback provided (Gan & Hattie, 2014). Similar to prompts, a pre-structured peer feedback form can improve the quality of the feedback given to peers (Gielen & De Wever, 2015). By providing students with carefully constructed rubrics and aggregating the assessments of four peers, their reliability and validity are even similar to those of teacher assessments (Cho, Schunn, & Wilson, 2006). In a setting that was less unidirectional and allowed some communication between feedback provider and

* Corresponding author. Maastricht University, Department of Educational Development & Research and Graduate School of Health Professions Education, P.O. Box 616, 6200 MD, Maastricht, the Netherlands.

E-mail address: kd.konings@maastrichtuniversity.nl (K.D. Könings).

<https://doi.org/10.1016/j.learninstruc.2018.11.007>

Received 30 January 2018; Received in revised form 9 November 2018; Accepted 25 November 2018

Available online 12 December 2018

0959-4752/ © 2018 Elsevier Ltd. All rights reserved.

receiver, it was demonstrated that assessors included more informative elaborations in the peer assessment when they received specific feedback requests from their peers beforehand (Voet, Gielen, Boelens, & de Wever, 2017). What remains unclear, however, is whether the effect of the aforementioned initiatives on assessors' learning persists when the support is no longer offered.

Next to the need to better understand the development of peer-assessment skills, the impact of task complexity on the quality of peer assessment warrants more attention. From the perspective of peer assessors, it was found that the quality of peer assessments is inversely related to the complexity of assessed tasks (Van Zundert, Sluijsmans, Könings, & Van Merriënboer, 2012). Moreover, when students perform peer assessments of tasks with increasingly higher levels of complexity, the quality of their peer assessments deteriorates more than their domain-specific task performance. Similarly, ability levels of students are known to impact the quality of peer assessments and of the feedback provided (Patchan et al., 2013), although earlier research has indicated that even young children can judge each other's strategies before they are able to use these adequately themselves (Siegler & Crowley, 1994). However, Patchan and colleagues have shown that when students must provide specific and high-quality feedback, which is more elaborate than giving a judgement, high-ability assessors provide better feedback on poor work of a peer than low-ability assessors do. Therefore, to improve the process whereby individual students learn peer-assessment skills, it is imperative that we consider the complexity of the task in relation to the quality of peer assessments. This is especially relevant when no predefined assessment criteria, rubrics or task solutions are available to assessors. In such cases students assess their peers' work based on their own understanding of the task or their own problem solution. We hypothesise that, in order to be able to assess their peers and provide specific feedback in these situations, students must be able to perform the task themselves.

Theoretical notions that may explain why the quality of peer assessments decreases as tasks become more complex are the cognitive load theory and the concept of learning hierarchies. According to cognitive load theory (Sweller, 1988, 2010; Sweller, Van Merriënboer, & Paas, 1998), the capacity of working memory – essential for the processing and storing of new information – is extremely limited. Students who are given complex tasks with many interacting elements must process many pieces of information simultaneously. Studying grammar as part of learning a language is an example of high element interactivity material, because the learner needs to consider many elements at once. To build sentences that are grammatically correct, the learner must attend to all the words within the sentence while also considering syntax, tense and verb endings. Vocabulary, by contrast, is an example of low element interactivity material. Although there are thousands of words to be learnt, most people can quickly learn some simple words because they can be learnt in isolation from all other words. The high demands on working memory that result from element interactivity (Paas, Renkl, & Sweller, 2003; Paas & Van Gog, 2006) may explain why performing peer assessments of more complex tasks is extra challenging: because of the complexity or higher element interactivity involved in performing the tasks, students have too limited cognitive capacity left for doing the peer assessments.

Another explanation as to why high cognitive load causes peer-assessment skills to deteriorate before doing the same to domain-specific skills is offered by the concept of 'learning hierarchies' (Van Zundert et al., 2012). The idea behind this notion is that proper peer assessment is *conditional* on a student's mastery of domain-specific skills (i.e. you cannot properly assess performance on a task if you cannot perform this task yourself), meaning that peer-assessment skills are positioned higher in the learning hierarchy than associated domain-specific skills (Anderson & Krathwohl, 2001; Bloom & Krathwohl, 1956). In case of simple tasks, students may be able to perform the task, compare their own performance with the assessee's and provide an accurate assessment. When tasks are more complex, however, students may have no

cognitive capacity left for comparing their own performance with that of the assessee, nor for providing an assessment, even though they may still be able to perform the task. The reason for this is that setting assessment criteria and applying them accordingly also becomes more difficult as task complexity increases (Tillema, Leenknecht, & Segers, 2011; Van Popta et al., 2017). In terms of Anderson and Krathwohl's (2001) taxonomy, if predefined criteria for the level 'evaluating' are missing, assessing their peer's performance is facilitated when assessors have already mastered lower-hierarchy skills, such as 'applying' or 'analysing'. Hence, if task complexity is high, cognitive overload might occur, causing peer assessment to suffer first. We therefore need instructional techniques that help students acquire peer-assessment skills and that reduce cognitive load when tasks are complex.

An instructional technique used to teach domain-specific skills and reduce cognitive load is *scaffolding* (Bruner, 1975; Greenfield, 1984, 1999; Rosenshine & Meister, 1992; Wood, Bruner, & Ross, 1976). Scaffolding refers to the process whereby the support given to students is gradually reduced to counteract the adverse effects of excessive task complexity. Full support is usually offered at the start of instruction and is then gradually withdrawn until the student is able to work independently. It has been shown that scaffolding of domain-specific skills reduces cognitive load and enhances learning (e.g., Renkl & Atkinson, 2003; Van Merriënboer, 1990; Van Merriënboer & De Croock, 1992; Van Merriënboer, Kirschner, & Kester, 2003). Scaffolding has also been found to be effective in improving students' use of each other's work during peer learning (Van Dijk & Lazonder, 2016). Similarly, the quality of peer assessments and feedback provided were shown to improve after students received scaffolding in the form of an assessment training module prior to the peer-assessment activities (Liu & Li, 2014). Ideally, however, students should receive support that is commensurate with their level of proficiency. This means that novices need full support or scaffolding to help them perform learning tasks (Van Merriënboer et al., 2003), because they have not yet developed cognitive schemas. As they progress, they do develop these schemas, reducing cognitive load and enabling them to perform the tasks more effectively.

Although there are many different types of scaffolding, the *completion strategy* is especially effective in scaffolding complex domain-specific tasks (Renkl & Atkinson, 2003; Renkl, Maier, Atkinson, & Staley, 2002; Stark, 2004; Van Merriënboer, 1990; Van Merriënboer & De Croock, 1992). Students first receive fully worked examples that include all the information that is necessary to understand and perform the tasks. As soon as students have acquired cognitive schemas, they receive guided assignments in which part of the necessary information has been omitted and which the students must complete. And finally, students receive conventional tasks without any tools or external support (Van Merriënboer & Kirschner, 2018), which requires them to draw on acquired peer-assessment skills for performing the peer assessment.

The purpose of the present study was to examine whether scaffolding by means of the completion strategy is as effective for teaching peer-assessment skills as it is for domain-specific skills. To this end, we addressed the following research questions: (1) what is the effect of scaffolding domain-specific learning tasks on domain-specific accuracy, time on task and perceived mental effort?; and (2) what is the effect of scaffolding of peer-assessment tasks on accuracy (both domain-specific and peer-assessment), time on task and perceived mental effort? Drawing from the results of earlier research, we hypothesised that the scaffolding of domain-specific skills would improve domain-specific accuracy, and reduce time on task and mental effort (Hypothesis 1). Along the same lines, we hypothesise that the scaffolding of peer-assessment skills would also improve peer-assessment accuracy, and reduce time on task and mental effort (Hypothesis 2). In a study using a 2x2 experimental design, students were offered only domain-specific scaffolding, only peer-assessment scaffolding, both types of scaffolding or no scaffolding at all.

Table 1
Explanation of the research steps.

Research step	Description
1. Observation	The researcher perceives a certain phenomenon and considers it for further research.
2. Problem statement	The researcher experiences this observation as a problem and formulates a problem statement.
3. Hypothesis	The researcher seeks to provide a logical explanation for the problem at hand.
4. Experimental stage	The researcher tests whether the hypothesis is correct or incorrect, by conducting an experiment and collecting data.
5. Results	The researcher displays the collected data synoptically, using graphs, diagrams or tables, for example.
6. Conclusion	The researcher compares the results with the hypothesis.

2. Method

2.1. Participants

Participants were 236 students from the first and third grades of two Dutch secondary schools that prepared students for either higher professional education in a 5-year programme (senior general education; $N_{\text{grade1}} = 49$, $N_{\text{grade3}} = 57$) or university in a 6-year programme (pre-university education; $N_{\text{grade1}} = 48$, $N_{\text{grade3}} = 82$). We looked for diversity in participants to ensure that there was variety in the level of perceived task complexity. In the electronic learning environment used in the study, students were randomly assigned to one of four conditions: no scaffolding ($N = 58$), peer-assessment scaffolding ($N = 59$), domain-specific scaffolding ($N = 60$), and both domain-specific and peer-assessment scaffolding ($N = 59$). Students had a mean age of 14.0 years ($SD = 1.18$) and 50.8% were female.

2.2. Materials

Students were presented with a sequence of six learning tasks and two test tasks in an electronic learning environment especially designed for this study. Tasks 1–3 were domain-specific tasks, whereas tasks 4–6 were a combination of domain-specific and peer-assessment tasks. Test tasks were similar to the combined learning tasks and measured both domain-specific and peer-assessment skills.

2.2.1. Content of tasks

The tasks contained a full description of a scientific experiment on topics such as operant conditioning, appetite and the effects of vitamin C. Students were asked to read these descriptions and to consequently identify the six different steps of doing scientific research which, for the purposes of our study, were defined as: (1) observation, (2) problem statement, (3) hypothesis, (4) experimental stage, (5) results and (6) conclusions. Table 1 provides a more detailed description of these steps.

To increase the complexity and cognitive load of the tasks, we used three principles of cognitive load theory. First, we added *additional elements* to render the descriptions more complex. For example, in the basic description of an experiment on appetite, food was prepared using varying levels of sunflower oil. In the more complex description, this was done using not only varying levels of oil but also varying levels of salt. Second, we added redundant information in the form of *irrelevant distractors*. For instance, we included irrelevant information about the occurrence of cardiomyopathy in dogs in a description of an experiment on the operant conditioning of a dog. Third, we raised complexity by adding *ambiguous relations* between elements. For example, the complex description of an experiment on gender differences and classroom behaviour would mention interactions between gender and communication patterns and relate these to IQ test scores, instead of only mentioning that both boys and girls scored similarly on IQ tests.

2.2.2. Learning tasks

Fig. 1 presents the order in which the learning tasks were offered to students, including the type of scaffolding they received in each

Fig. 1. Domain-specific scaffolding and peer-assessment scaffolding of learning tasks in the four conditions.

condition. To complete learning tasks 1–3, students in all conditions needed to apply their domain-specific skills, whereas for tasks 4–6 they needed a combination of domain-specific and peer-assessment skills. After the first three tasks, we gave participants a 5-min break during which they watched a video clip about a monkey in a funny research setting. We included this break to support students' learning, as numerous studies have indicated that providing breaks during study time has beneficial effects on learning (e.g., Ariga & Lleras, 2011). By allowing students a short rest, we hoped to help them restore the attentional resources needed to invest mental effort in the upcoming tasks (Marranges, Schmeichel, & Baumeister, 2017).

After the break, participants performed the remaining three learning tasks that called upon both their domain-specific and peer-assessment skills. The domain-specific instructions were similar to those of the first three tasks. The peer-assessment tasks required students to judge a fictitious peer's performance in terms of accuracy on a task identical to the domain-specific tasks (for which students needed to identify the six research steps in an experiment description). All students were given the same answers from a fictitious peer, designed by the researchers. Students individually evaluated these answers and judged their correctness.

As a *measure* of students' accuracy in the learning phase, we collected and scored their answers to the open-ended questions of learning tasks 3 and 6; these were the only tasks that were identical in all conditions. For each of the six steps that were correctly identified, we gave students a score of '1'. Otherwise, a score of '0' was assigned. As to the peer-assessment parts of learning task 6, we gave students a score of '1' for each correct judgement and a score of '0' in all other cases. We subsequently computed mean scores per task.

2.2.3. Scaffolding

Depending on the condition, students received scaffolding of domain-specific skills only, of peer-assessment skills only, of both types of skills or no scaffolding at all. In the conditions with *domain-specific scaffolding*, students performed three tasks with gradually diminishing support: task 1 was a worked example (WE), task 2 was a guided assignment (GA; or completion strategy, as mentioned in the Introduction) and task 3 contained open-ended questions (OQ) only. In the worked examples the six research steps were identified and linked to the relevant parts of the experiment description. Tasks with a guided assignment were similar to the worked examples, but now students were asked to enter the correct steps into a blank space. The open-ended questions asked students for each step: 'Which step is this and can you explain your answer?' In the conditions without domain-specific scaffolding, only open-ended questions were used.

In the conditions with *peer-assessment scaffolding*, students also performed three tasks with gradually diminishing support: worked examples, guided assignments and open-ended questions. The worked example presented a fictitious teacher's assessment of the accuracy of a fictitious peer's answers, meaning that this teacher judged the correctness of the peer's answer and gave a brief explanation for this judgement. The guided assignments resembled the worked examples, but now the teacher's judgement was omitted and had to be entered into a blank space by the student. We used teachers' assessments in the scaffolding, because students more readily view teachers as authoritative and reliable sources of information, whereas they are more likely to question a peer. The open-ended questions asked: 'Assess the accuracy of your classmate's answer and explain your judgement'. In the conditions without peer-assessment scaffolding, students were given three tasks with only open-ended questions.

Fig. 2 presents a fragment of a combined learning task with peer-assessment scaffolding. In the upper text box, all six research steps are presented. In the lower box, one research step is presented at a time. The left box, labelled 'teacher', contains domain-specific instructions about the research step. The right box, labelled 'classmate', presents the answer of a fictitious peer (e.g. 'This is a hypothesis. The hypothesis is

that coldness narrows the blood vessels in the upper part of the respiratory system and therefore the immune system functions less adequately'). The lower box, labelled 'assess whether your classmate has correctly identified the research step', contains instructions on how to assess whether the fictitious peer gave a correct answer. In case of open-ended questions, students were asked to indicate if the peer's answer was correct, for example: 'No, I don't think this is a hypothesis'.

2.2.4. Test tasks

Two test tasks assessed both domain-specific and peer-assessment skills. They were similar to the combined learning tasks, but the topics of the described experiments were different. Open-ended questions of domain-specific test tasks asked students to identify five of the six research steps, because after providing five answers, only one answer option would remain available for the sixth question. The steps were presented in a random order to prevent participants from guessing and, additionally, were shown in subsequent screens to ensure that the questions were answered independently of each other and were not affected by answers given before. Accordingly, the possibility to return to the previous screen in the electronic learning environment was disabled. Two peer-assessment test tasks measured students' ability to judge a *peer's* knowledge and understanding of the research steps. Students received the answers to the domain-specific tasks given by a fictitious peer and were asked to indicate for each answer whether it was correct or not and why.

Accuracy *measures* of domain-specific task performance were computed by giving students a score of '1' for each step that was correctly identified and a score of '0' for each incorrect answer. We then computed a mean score of the two test tasks, which varied between '0' and '1'. In the case of peer-assessment tasks, we gave students a score of '1' for each correct judgement and a score of '0' in all other cases. Mean scores again varied between 0 and 1. All answers to test tasks were coded by two independent raters (i.e. the second author and a student assistant). The correlation between the scores of both raters was $r = 0.93$ in the case of domain-specific tasks and $r = 0.81$ in the case of peer-assessment tasks. We subsequently used the scores of the second author.

2.2.5. Time on task

The time students spent working on the learning tasks and test tasks was logged, separately for domain-specific tasks and peer-assessment tasks, into the electronic learning environment.

2.2.6. Cognitive load

Since cognitive load is typically operationalised in terms of perceived mental effort (Paas, Tuovinen, Tabbers, & Van Gerven, 2003), we used Paas' subjective 9-point mental effort rating scale (1992) to measure cognitive load. Students were asked to answer the question: 'How much effort did it take to study/perform this task?' on a scale from 1 'very, very little effort' to 9 'a great deal of effort'. In our study, the scale's Cronbach's alpha was .87, which is similar to the alpha of .90 reported in Paas' original article (1992).

2.3. Procedure

Students worked individually at a computer and performed all tasks at their own pace. Before students performed the first three domain-specific learning tasks, they were asked to answer questions about age and gender. After the break of approximately 5 min, we gave them the remaining learning tasks with combined instructions and the two test tasks immediately after that. As soon as students had completed each learning or test task, they rated their perceived mental effort. The entire procedure took approximately an hour and a half. All participants received a €7.50 gift voucher for their participation.

Fig. 2. Example of a learning task fragment with peer-assessment scaffolding and no domain-specific scaffolding.

the posttest. We included the following covariates: the level of education (senior general secondary education or pre-university education) and the student's grade (first and third grade). Additionally, we performed univariate ANCOVAs to explore if and how the individual variables differed across conditions. When we found interactions, we ran post-hoc tests with Bonferroni correction to compare outcomes across conditions. As a measure of effect size we used eta-squared (η^2), where 0.01 corresponded to a small effect, 0.06 to a medium effect and 0.14 to a large effect (Cohen, 1988). There was no multicollinearity between the measures used.

Table 2
Means (Standard Deviations in parentheses) of accuracy, time on task and mental effort on learning tasks 3 and 6 and the test tasks for each condition.

	Domain-specific scaffolding (N = 60)			Peer-assessment scaffolding (N = 59)			Domain-specific and peer-assessment scaffolding (N = 59)			No scaffolding (N = 58)		
	Task 3	Task 6	Test	Task 3	Task 6	Test	Task 3	Task 6	Test	Task 3	Task 6	Test
Domain-specific accuracy ^a	0.35 (0.26)	0.39 (0.26)	0.31 (0.20)	0.29 (0.25)	0.38 (0.27)	0.21 (0.16)	0.28 (0.23)	0.34 (0.25)	0.19 (0.16)	0.31 (0.20)	0.37 (0.23)	0.23 (0.16)
Peer-assessment accuracy ^a	–	0.37 (0.27)	0.39 (0.19)	–	0.54 (0.27)	0.38 (0.21)	–	0.57 (0.22)	0.39 (0.20)	–	0.49 (0.25)	0.35 (0.18)
Domain-specific time on task ^b	761.25 (266.10)	409.29 (143.54)	286.33 (106.27)	561.60 (250.00)	401.78 (128.57)	255.96 (90.91)	748.20 (267.28)	416.14 (125.14)	270.84 (86.29)	612.45 (220.08)	377.86 (129.52)	271.58 (92.65)
Peer-assessment time on task ^b	–	365.66 (121.54)	276.70 (84.97)	–	392.78 (133.87)	264.88 (75.23)	–	392.40 (120.66)	265.54 (77.88)	–	368.00 (124.30)	270.13 (79.35)
Mental effort ^c	5.36 (1.54)	4.77 (1.70)	5.39 (1.68)	5.32 (1.48)	5.50 (1.62)	5.77 (1.67)	5.81 (1.28)	5.51 (1.56)	6.09 (1.42)	5.41 (1.52)	5.02 (1.70)	5.60 (1.72)

Note. No peer assessment in learning task 3.

^a Scale = 0–1.

^b Time in s.

^c Scale = 0–9.

Due to technical problems in the electronic learning environment, we had 24 cases with missing data on time on task and 32 cases with missing data on mental effort. For time on the domain-specific test task and the peer-assessment test task, these missing data were randomly distributed across conditions (respectively, $\chi^2(3, 236) = 4.63, p = .20$; $\chi^2(3, 236) = 4.43, p = .22$). This was also the case for the mental effort expended on the test tasks ($\chi^2(3, 236) = 2.66, p = .45$). Consequently, we deleted these cases listwise from the relevant analyses. We checked skewness and kurtosis for all variables: skewness values were all in an acceptable range (between -0.10 and 0.97), but kurtosis varied between -0.65 and 0.53 with one value of 4.76 for time on learning task 6 due to a single outlier. After we excluded this outlier, kurtosis decreased to 0.19 . We decided to exclude the outlier from all further analyses regarding time on task. Table 2 gives an overview of the means and standard deviations for accuracy, time on task and mental effort on learning and test tasks for each condition.

Regarding the results on *accuracy*, the MANCOVA showed no main effect of domain-specific scaffolding, $V = 0.02, F(1, 226) = 1.13, p = .35$ (Hypothesis 1). We did find an overall significant effect of peer-assessment scaffolding on students' accuracy on the tasks, $V = 0.16, F(1, 226) = 8.75, p < .01$ (Hypothesis 2). Students receiving peer-assessment scaffolding outperformed their peers who had not received this type of scaffolding in terms of peer-assessment accuracy on learning phase task 6, $F(1, 230) = 16.27, p < .01, \eta^2 = 0.06$. Yet, we found no main effect of peer-assessment scaffolding on accuracy during the peer-assessment test tasks, $F(1, 230) = 0.25, p = .62, \eta^2 = 0.001$. Additionally, we found that students who had received peer-assessment scaffolding performed worse on domain-specific test tasks than their peers without this type of scaffolding, $F(1, 230) = 11.66, p < .01, \eta^2 = 0.04$.

The combination of domain-specific and peer-assessment scaffolding impacted accuracy, $V = 0.08, F(1, 226) = 3.77, p < .01$, as shown by significant interaction effects on peer-assessment accuracy on learning task 6, $F(1, 230) = 5.48, p = .02, \eta^2 = 0.02$ and on domain-specific accuracy on the test tasks, $F(1, 230) = 4.75, p = .03, \eta^2 = 0.02$. On learning task 6, peer-assessment accuracy was highest for those students receiving both types of scaffolding, followed by peer-assessment scaffolding and no scaffolding, and it was lowest for the group receiving domain-specific scaffolding only. On the domain-specific test tasks, receiving domain-specific scaffolding only yielded the highest accuracy scores ($M = 0.31, SD = 0.20$), as visualised in Fig. 3; the groups receiving both types of scaffolding ($M = 0.19, SD = 0.16$) or peer-assessment scaffolding only ($M = 0.21, SD = 0.16$) did not show higher accuracy than the no-scaffolding group ($M = 0.23, SD = 0.16$). The interaction effect on peer-assessment accuracy at the test tasks was non-significant, $F(1, 230) = 0.62, p = .43, \eta^2 < 0.01$ (see Fig. 4).

As for *time on task*, we found an effect of domain-specific scaffolding, $V = 0.16, F(1, 199) = 7.73, p < .01$ (Hypothesis 1). Students who had received domain-specific scaffolding spent more time on tasks, especially on learning task 3, than their peers, $F(1, 203) = 28.52, p < .01, \eta^2 = 0.12$. There was no significant main effect of peer-assessment scaffolding, $V = 0.04, F(1, 199) = 1.73, p = .13$ (Hypothesis 2), nor did we find any interaction effect, $V = 0.02, F(1, 199) = 0.64, p = .67$.

With respect to perceived *mental effort*, we found no main effect of domain-specific scaffolding, $V = 0.01, F(1, 196) = 0.81, p = .49$ (Hypothesis 1). However, we did find an effect of peer-assessment scaffolding on mental effort, $V = 0.05, F(1, 196) = 3.24, p = .02$ (Hypothesis 2). On learning task 6, students who had received peer-assessment scaffolding reported higher mental effort than those who had not, $F(1, 198) = 8.79, p < .01, \eta^2 = 0.04$. No interaction effect was found, $V = 0.02, F(1, 196) = 1.08, p = .36$.

Since our results did not reveal a match between mental effort and accuracy, we checked whether the relation between mental effort and accuracy yielded an appropriate theoretical framing for the effects of scaffolding. As expected, mental effort and accuracy were negatively

Fig. 3. Interaction effect of domain-specific and peer-assessment scaffolding on domain-specific accuracy at the test task.

Fig. 4. Effect of domain-specific and peer-assessment scaffolding on peer-assessment accuracy at the test task.

and mostly significantly related in the following tasks: learning task 3, $r = -0.14$, $p = .03$; learning task 6_{domain-specific}, $r = -0.16$, $p = .02$; learning task 6_{peer assessment}, $r = -0.09$, $p = .16$; test tasks_{domain-specific} $r = -0.22$, $p < .01$; and test tasks_{peer assessment} $r = -0.17$, $p = .02$. These findings indicate that using accuracy and mental effort, as an operationalisation of cognitive load, was appropriate in this study.

4. Discussion

This study investigated whether scaffolding of peer-assessment skills positively affects accuracy, time on task and perceived mental effort in the same way that scaffolding of domain-specific skills does. Our results partly support the first hypothesis that scaffolding of domain-specific skills improves domain-specific accuracy and reduces both time on task and perceived mental effort. Domain-specific scaffolding yielded highest accuracy on the domain-specific test tasks, unless it was combined with peer-assessment scaffolding. This is in line with results of previous research which demonstrated that domain-specific scaffolding improves performance (e.g., Renkl & Atkinson, 2003; Van Merriënboer, 1990; Van Merriënboer & De Croock, 1992; Van Merriënboer et al., 2003). Domain-specific scaffolding did not reduce perceived mental effort and increased time-on-task during the learning phase, but not in the test phase. Regarding this absence of effects in the test phase, it should be noted that previous interventions generally lasted longer than our experiment. We only included domain-specific scaffolding in three of the six learning tasks. Enabling students to practise on a larger number of tasks might make scaffolding more powerful and strengthen its impact on both accuracy and time on task.

Our results partly support the second hypothesis that peer-assessment scaffolding improves peer-assessment accuracy and reduces time on task and perceived mental effort. However, we also obtained some unexpected results. First, peer-assessment scaffolding increased peer-assessment accuracy during learning, but increased rather than decreased perceived mental effort. Second, peer-assessment accuracy during learning was highest when peer-assessment scaffolding was offered alongside domain-specific scaffolding. The same did not hold true for the test tasks, where peer-assessment scaffolding did not affect peer-assessment accuracy. Third, and most surprisingly, peer-assessment scaffolding had a *negative* impact on domain-specific accuracy at the test. More specifically, offering peer-assessment scaffolding either with or without domain-specific scaffolding did not yield any higher accuracy than no scaffolding at all (i.e. it is best to use domain-specific scaffolding *only*). The reason for this small albeit surprising, negative effect of peer-assessment scaffolding on domain-specific accuracy may be that peer-assessment scaffolding interferes with students' yet incomplete domain-specific cognitive schemas. As can be inferred from Table 2, students did not perform very well on domain-specific accuracy during the learning phase, suggesting that they did not yet have sufficient domain-specific knowledge the moment peer-assessment scaffolding was introduced. In other words, peer-assessment scaffolding was introduced at too early a stage relative to students' domain-specific knowledge. Indeed, as the Process Model for Providing Online Peer Feedback emphasises (Van Popta et al., 2017), peer assessment requires additional cognitive processes, such as evaluating, reflecting, regulating own thinking, critical thinking, explaining and taking different perspectives. Our results suggest that scaffolding the process whereby students assess their peers when they have not yet fully mastered the task at hand can have disturbing effects on students' learning of the task content.

This raises the question of why peer-assessment scaffolding had no effect on peer-assessment accuracy at the test. Or, why did it only negatively affect domain-specific accuracy at the test, and not peer-assessment accuracy? A plausible answer is that a negative effect could not be found because students did not score above chance level on the peer-assessment test, thus, producing a clear bottom effect. For each of the five questions students were asked to indicate whether the answer

of the fictitious peer was correct or not, yielding a chance score of 0.5. On the peer-assessment test, students in all four conditions scored around this chance level. Even after we excluded the worst performing students (i.e. those who had answered less than 10% or 20% of questions correctly on learning task 3 or the peer-assessment test task) from the analyses to see whether their scores impacted the findings regarding accuracy, time on task and mental effort, the results relevant to our hypotheses remained unaltered. Hence, despite the robustness of findings, we deem the bottom effect a potential culprit. Nevertheless, we did find that peer-assessment scaffolding had positive effects on peer-assessment accuracy during the learning phase, which finding ties in nicely with the research alluded to in the Introduction demonstrating that specific tools such as prompts, pre-structured peer feedback forms or feedback requests can assist students in assessing their peers. The fact that we did not find effects in the test phase, however, underscores the need for further research into the long-term effects of tools on the acquisition of peer-assessment skills, for instance by testing the effects on assessors when tools are no longer available.

A theoretical implication of this study is that the learning hierarchy principles indeed apply when scaffolding peer-assessment and domain-specific skills. We expected domain-specific skills to be a prerequisite for learning peer-assessment skills, since the latter are positioned higher up in the learning hierarchy (Anderson & Krathwohl, 2001). Students must be able to perform the task themselves before starting the cognitive processes of reviewing the peer product and providing peer feedback (Van Popta et al., 2017). This explained why an increase in task complexity adversely affected peer-assessment performance before it negatively affected domain-specific performance (Van Zundert et al., 2012). At first sight, it may therefore seem strange that the present study found peer-assessment scaffolding to have negative effects on domain-specific accuracy, but not on peer-assessment accuracy. Yet, taking into account the bottom effect of the peer-assessment test may help to clarify this result. If the cognitive processes necessary for learning to review and assess a peer product are very demanding, the cognitive processes involved in performing the domain-specific task can be disturbed. This suggests that domain-specific knowledge is a prerequisite for being able to learn to perform peer assessment. In the literature on supporting peer assessment and also self-assessment, however, little attention has been paid to the relation with domain-specific skills. Our study shows that, when developing models for peer assessment and self-assessment, it is important to explicitly factor in students' prior knowledge and their development of domain-specific knowledge.

Hence, endorsing learning hierarchy principles and the Process Model for Providing Online Peer Feedback, our results suggest that scaffolding of peer-assessment skills will only be effective when domain-specific skills are sufficiently developed, for example through the use of domain-specific scaffolding. We invite future researchers to investigate this idea further and to study what happens when scaffolding strategies are implemented in longitudinal training programmes or curricula. We propose that domain-specific scaffolding should be used until students are able to perform domain-specific skills independently and that *only then* peer-assessment scaffolding should be introduced. Furthermore, our findings only allow us to draw conclusions on the effects of peer-assessment scaffolding, not on the effects of using peer assessment tasks per se. Future studies would benefit from an additional condition in which students do not perform peer-assessment tasks.

Methodologically, a few limitations need consideration. First, we did not test whether the eight tasks (i.e., experiment descriptions) were of equal complexity. Although failure to do so did not impact outcomes since tasks were identical across conditions, ensuring similar complexity would have optimised the design. Second, we increased task complexity by adding extraneous cognitive load (i.e. load that is extraneous to the task because it does not directly contribute to performance or learning; Sweller et al., 1998) rather than intrinsic cognitive load, which would be an interesting idea for future research.

Additionally, since students' perceived mental effort was at a moderate level, the tasks may not have been sufficiently complex. The experimental set-up might have been more powerful in inducing effects of scaffolding if tasks were even more complex or students were less advanced. Furthermore, scaffolding was identical for all students and not adaptive to their level of comprehension and skill acquisition. We therefore recommend that future studies design scaffolding support appropriate to students' needs.

Another limitation of the current study is that the circumstances under which students assessed their peers were only a remote reflection of reality. The assessment concerned a brief experiment only and took place anonymously in an electronic learning environment, and the peer-assessment task was small and simplified compared to a more natural setting. It is not unusual that students are required to give feedback to peers while they have not completely mastered the subject matter themselves. This fits the use of current educational methods such as self-directed learning (e.g., Corbalan, Kester, & Van Merriënboer, 2008; Kauffman, Zhao, & Yang, 2011), where students take ownership of their learning (Könings, Seidel, & Van Merriënboer, 2014), and collaborative learning (e.g., Saab, Van Joolingen, & Van Hout-Wolters, 2005; Schmitz & Winskel, 2008). In an authentic context, however, peer assessments are likely to form an integral part of the curriculum, involving social and additional cognitive factors. Moreover, it could be that rather than being anonymous, peer assessments are communicated to the peers involved. Finally, although peer assessments usually encompass several aspects, students in our study were only required to judge the performance of a peer in terms of accuracy. In doing so, we largely neglected other skills that, according to Sluijsmans (2002), are important in peer assessment, specifically: defining assessment criteria and providing feedback for future learning. It is therefore imperative that future studies consider the presence and effects of additional social and cognitive factors as well as the conditions under which peer-assessment scaffolding does or does not enhance student learning.

The practical implications of our study are straightforward. The results reaffirm that domain-specific scaffolding is a proven, powerful strategy to foster the development of domain-specific skills. In contrast, for students with weak domain-specific skills the scaffolding of peer-assessment skills is a barrier rather than a boost to learning. This is especially true when it is coupled with domain-specific scaffolding. Arguably, peer-assessment scaffolding should only be introduced *after* students have mastered domain-specific skills. Yet, we believe that it remains important to address peer assessment in education to prepare students for tomorrow's workplaces. In doing so, however, educationalists should be aware of the risks inherent in implementing peer-assessment scaffolding at too early a stage in learning environments with complex tasks. Peer-assessment tasks must be carefully designed, be appropriately complex and be tailored to students' level of domain-specific skills.

Note

For additional information about experimental materials, researchers can contact the corresponding author.

Acknowledgements

This research project is funded by the Netherlands Organisation for Scientific Research under project number 411-05-110.

References

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy of learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. New York: Longman.
- Ariga, A., & Lleras, A. (2011). Brief and rare mental 'breaks' keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition*, 118, 439–443.
- Ashenafi, M. M. (2017). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education*, 42, 226–251.
- Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain*. New York: Longmans.
- Bruner, J. S. (1975). From communication to language: A psychological perspective. *Cognition*, 3, 255–287.
- Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education*, 25, 78–84.
- 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, 891–901.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.
- Corbalan, G., Kester, L., & Van Merriënboer, J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on efficiency and task involvement. *Contemporary Educational Psychology*, 33, 733–756.
- Demiraslan Çevik, Y., Haşlamam, T., & Çelik, S. (2015). The effect of peer assessment on problem solving skills of prospective teachers supported by online learning activities. *Studies In Educational Evaluation*, 44, 23–35.
- 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, 861–878.
- Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, 52, 315–325.
- Gielen, S., Tops, L., Dochy, F., Onghena, P., & Smeets, S. (2010). A comparative study of peer and teacher feedback forms in secondary school writing curriculum. *British Educational Research Journal*, 36, 143–162.
- Greenfield, P. M. (1984). A theory of the teacher in the learning activities of everyday life. In B. Rogoff, & J. Lave (Eds.). *Everyday cognition: Its development in social context*. Cambridge, MA: Harvard University Press.
- Greenfield, P. M. (1999). Historical change and cognitive change: A two-decade follow-up study in Zinacantan, a Maya community in Chiapas, Mexico. *Mind, Culture and Activity*, 6, 92–98.
- Kauffman, D. F., Zhao, R., & Yang, Y. S. (2011). Effects of online note taking formats and self-monitoring prompts on learning from online text: Using technology to enhance self-regulated learning. *Contemporary Educational Psychology*, 36, 313–322.
- Könings, K. D., Seidel, T., & Van Merriënboer, J. J. G. (2014). Participatory design of learning environments: Integrating perspectives of students, teachers, and designers. *Instructional Science*, 42, 1–9.
- Liu, X., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education*, 39, 275–292.
- Maranges, H. M., Schmeichel, B. J., & Baumeister, R. F. (2017). Comparing cognitive load and self-regulatory depletion: Effects on emotions and cognitions. *Learning and Instruction*, 51, 74–84.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84, 429–434.
- Paas, F., Renkl, A., & Sweller, J. (2003a). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38, 1–4.
- Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38, 63–71.
- Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, 16, 87–91.
- Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science*, 41, 381–405.
- Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, 38, 15–22.
- Renkl, A., Maier, U. H., Atkinson, R. K., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education*, 70, 293–315.
- Rosenshine, B., & Meister, C. (1992). The use of scaffolds for teaching higher-level cognitive Strategies. *Educational Leadership*, 49, 26–33.
- Saab, N., Van Joolingen, W. R., & Van Hout-Wolters, B. H. A. M. (2005). Communication in collaborative discovery learning. *British Journal of Educational Psychology*, 75, 603–621.
- Schmitz, M. J., & Winskel, H. (2008). Towards effective partnership in a collaborative problem-solving task. *British Journal of Educational Psychology*, 78, 581–596.
- Siegler, R. S., & Crowley, K. (1994). Constraints on learning in nonprivileged domains. *Cognitive Psychology*, 27, 194–226.
- Sluijsmans, D. M. A. (2002). *Student involvement in assessment: The training of peer assessment skills* Unpublished doctoral dissertation. Heerlen, The Netherlands: Open University of The Netherlands.
- Stark, R. (2004). Implementing example-based learning and teaching in the context of vocational school education in business administration. *Learning Environments Research*, 7, 143–163.
- Strijbos, J. W., Narciss, S., & Dünnebie, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, 20, 291–303.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive

- load. *Educational Psychology Review*, 22, 123–138.
- Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design: Recent developments. *Educational Psychology Review*, 10, 251–296.
- Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning - a review of research studies. *Studies In Educational Evaluation*, 37, 25–34.
- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68, 249–276.
- Topping, K. J. (2009). Peer assessment. *Theory Into Practice*, 48, 20–27.
- Van Dijk, A. M., & Lazonder, A. W. (2016). Scaffolding students' use of learner-generated content in a technology-enhanced inquiry learning environment. *Interactive Learning Environments*, 24, 194–204.
- Van Merriënboer, J. J. G. (1990). Strategies for programming instructions in high school: Program completion vs. program generation. *Journal of Educational Computing Research*, 6, 265–285.
- Van Merriënboer, J. J. G., & De Croock, M. B. M. (1992). Strategies for computer-based programming instruction: Program completion vs. program generation. *Journal of Educational Computing Research*, 6, 365–394.
- Van Merriënboer, J. J. G., & Kirschner, P. A. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design* (3rd Rev. Ed). New York: Routledge.
- Van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load of a learner's mind: Instructional design for complex learning. *Educational Psychologist*, 38, 5–13.
- Van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P. R. J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review*, 20, 24–34.
- 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*, 32, 137–145.
- Van Zundert, M. J., Sluijsmans, D. M. A., & Van Merriënboer, J. J. G. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20, 270–279.
- Voet, M., Gielen, M., Boelens, R., & de Wever, B. (2017). Using feedback requests to actively involve assesseees in peer assessment: Effects on the assessor's feedback content and assessee's agreement with feedback. *European Journal of Psychology of Education*, 33, 145–164.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17, 89–100.