Instructional Science (2020) 48:651–697
https://doi.org/10.1007/s11251-020-09525-2

ORIGINAL RESEARCH

When failure fails to be productive: probing the effectiveness of productive failure for learning beyond STEM domains

Valentina Nachtigall[1] · **Katja Serova**[1] · **Nikol Rummel**[1]

Received: 21 November 2018 / Accepted: 26 October 2020 / Published online: 19 November 2020

Abstract

The current work builds on research demonstrating the effectiveness of Productive Failure (PF) for learning. While the effectiveness of PF has been demonstrated for STEM learning, it has not yet been investigated whether PF is also beneficial for learning in non-STEM domains. Given this need to test PF for learning in domains other than mathematics or science, and the assumption that features embodied in a PF design are domain-independent, we investigated the effect of PF on learning social science research methods. We conducted two quasi-experimental studies with 212 and 152 10th graders. Following the paradigm of typical PF studies, we implemented two conditions: PF, in which students try to solve a complex problem prior to instruction, and Direct Instruction (DI), in which students first receive instruction followed by problem solving. In PF, students usually learn from their failure. Failing to solve a complex problem is assumed to prepare students for deeper learning from subsequent instruction. In DI, students usually learn through practice. Practicing and applying a given problem-solving procedure is assumed to help students to learn from previous instruction. In contrast to several studies demonstrating beneficial effects of PF on learning mathematics and science, in the present two studies, PF students did not outperform DI students on learning social science research methods. Thus, the findings did not replicate the PF effect on learning in a non-STEM domain. The results are discussed in light of mechanisms assumed to underlie the benefits of PF.

Keywords Productive failure · Direct instruction · Time of instruction · Problem solving prior to instruction

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11251-020-09525-2) contains supplementary material, which is available to authorized users.

✉ Valentina Nachtigall
valentina.nachtigall@rub.de

Katja Serova
katja.serova@rub.de

Nikol Rummel
nikol.rummel@rub.de

[1] Institute of Educational Research, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany

Introduction

In the last decade there has been increased interest in investigating the effectiveness of approaches with problem solving prior to instruction, such as Productive Failure (PF), for students' knowledge acquisition. Several recent reviews of studies testing for the PF effect (cf., Kapur 2015; Loibl et al. 2017; Darabi et al. 2018) have demonstrated beneficial effects of PF (or problem-solving prior to instruction approaches more generally) on students' acquisition of conceptual knowledge. This increased interest in the effectiveness of PF might be partly due to the expectation that PF holds the potential to resolve the assistance dilemma (Kapur and Rummel 2009). The assistance dilemma (Koedinger and Aleven 2007) describes a critical and hitherto unanswered question in instructional science: How can the giving and withholding of instructional support be balanced in order to promote students' learning most effectively?

The PF approach suggests temporarily withholding instructional support in order to foster students' development of a deep conceptual understanding. Specifically, PF combines two successive learning phases, an initial problem-solving phase and a subsequent instruction phase. During the former, students collaboratively generate solution ideas for a complex and novel problem without instructional support (Kapur and Bielaczyc 2012). During the latter, the instructor builds upon erroneous student solutions and compares and contrasts the features of these erroneous solutions with the components of the canonical solution (Kapur and Bielaczyc 2012). Most students fail to solve the complex problem canonically during the initial problem-solving phase, i.e. they generate incomplete or erroneous solution ideas due to lacking the specific knowledge. However, this initial failure is assumed to productively prepare students for learning during subsequent instruction (e.g. Kapur 2016). When students fail to solve the problem (often expressed in erroneous solution ideas) or at least when confronted with their erroneous solution ideas during instruction, they should experience the limitations of their existing understanding (e.g. Loibl and Rummel 2014a). These experiences of failure are assumed to challenge students' current understanding and thus prepare them for revising their current understanding (cf., Tawfik et al. 2015).

The effectiveness of PF for students' acquisition of conceptual knowledge has predominantly been examined and replicated for learning in STEM domains, especially mathematics. Thus, there is a lack of empirical evidence regarding whether the PF effect depends on the structuredness of the domain or whether it also occurs for learning in less structured domains (Kapur 2015; Loibl et al. 2017). Accordingly, Kapur (2015) and Loibl et al. (2017) called for research investigating the effectiveness of PF for student learning in domains other than mathematics or science. To address this research gap, the present work attempts to probe the effectiveness of PF for learning beyond STEM domains. Specifically, we examine the effectiveness of PF for students' acquisition of knowledge about social science research methods in two quasi-experimental studies.

In the following sections, we first describe the features of the PF design and the reasons why these features are hypothesized to be important for the effectiveness of PF. Next, we discuss potential particularities of STEM versus non-STEM domains and how these may be linked to the effectiveness of PF. Finally, we give an overview of previous research on PF, focusing on the design features implemented in these studies and the domains that they investigated. The description of the PF design features, the discussion of potential requirements that should be met by the domain in a PF setting, and the literature overview lead to several criteria that guided the design of our two studies.

Theoretical background

PF stresses the benefits of failure, caused by temporarily withholding certain support structures (i.e., structure of a problem-solving task and instructional support from a teacher) during a problem-solving activity, for students' learning (Kapur 2008). The idea of PF builds strongly on VanLehn's (1988) theory of impasse-driven learning, postulating that learning only occurs when learners' knowledge is incomplete, such that they reach an impasse during problem-solving. According to research on PF, the effectiveness of these impasses or failures for learning is determined by particular features of the instructional design and specific learning-related mechanisms, which are described in the following section.

Productive Failure: design features and learning mechanisms

Kapur (2015) claims that the PF effect depends on particular features of the instructional design. The review by Loibl et al. (2017) also suggests that the effectiveness of PF (or problem solving prior to instruction more generally) relates to specific features of the instructional design. These features are expected to trigger the following three learning-related mechanisms which are hypothesized to underlie the effectiveness of PF (Loibl et al. 2017): (1) activation of prior knowledge, (2) awareness of knowledge limitations, and (3) recognition of the deep features of the targeted learning concepts. The instructional design as well as the instructor have to promote these processes in particular ways (for an overview, see Table 1). Specifically, asking students to collaboratively invent solution ideas to a complex and novel problem prior to instruction is expected to trigger mechanisms (1) and (2). For this purpose, the selected problem has to meet a certain level of complexity. Comparing and contrasting features of erroneous solution ideas with components of the canonical solution during subsequent instruction is assumed to promote mechanisms (2) and (3). For this purpose, the instruction has to use students' erroneous solution ideas.

These design features and underlying mechanisms of PF demonstrate that PF is both a *learning activity* that triggers certain learning mechanisms in learners and an *instructional approach*, as the instructor and the instructional design must foster these mechanisms in particular ways. In the following paragraphs, we describe the three mechanisms in more detail.

Activation of prior knowledge

The first mechanism (i.e., activation of prior knowledge) is expected to emerge during the initial problem-solving phase of PF. When students generate intuitive solution ideas to a complex and novel problem prior to instruction, PF is hypothesized to help them activate relevant prior knowledge (Kapur 2016; Loibl et al. 2017; Newman and DeCaro 2018). In turn, this should prepare students for "future learning" (Schwartz and Martin 2004), that is, for learning the targeted concepts during subsequent instruction. According to schema theory, the activation of preexisting schemas (i.e., prior knowledge) is a crucial requirement for the integration of new knowledge and thus for the adaptation of one's preexisting schemas (Sweller 1988; Sweller et al. 1998).

In a PF setting, the process of prior knowledge activation is expected to depend on the design of the problem that students are asked to solve. According to Kapur and Bielaczyc

Table 1 Overview of design features and mechanisms hypothesized to underlie the PF effect

	Initial problem-solving phase	Delayed instruction phase
Learning mechanisms	Activation of prior knowledge Awareness of knowledge limitations	Awareness of knowledge limitations Recognition of deep features
Design features	Asking students to collaboratively invent solution ideas to a complex and novel problem Problem meets a sweet spot of complexity, i.e., addresses students' prior knowledge related to the targeted learning concept	Comparing and contrasting features of erroneous solution ideas with components of canonical solutions Using students' erroneous solution ideas

(2012), the problem should address students' prior knowledge by meeting a sweet spot of complexity or, as Jacobson et al. (2015) conjecture, by meeting a zone of proximal failure. That is, students should not be able to develop the canonical solution during the initial problem-solving phase. At the same time, they should be able to invent intuitive ideas without becoming frustrated with the difficulty. Previous PF studies have used students' invented solution ideas as a measure for PF students' activation of prior knowledge (e.g. Kapur and Bielaczyc 2012; Loibl and Rummel 2014b). As PF students usually generate their intuitive solution ideas in small groups and without any instructional support during the initial problem-solving phase, they can build only on their prior knowledge in order to invent solution ideas (cf., Kapur 2016). These invented solution ideas therefore reflect PF students' prior knowledge (Kapur 2012; Loibl and Rummel 2014b).

Awareness of knowledge limitations

The second mechanism (i.e., awareness of knowledge limitations) is expected to emerge during the delayed instruction phase. By comparing and discussing erroneous student solutions during the delayed instruction phase, PF is hypothesized to help students become aware of the limitations and gaps in their prior knowledge (Loibl et al. 2017). The process of helping learners to become aware of their knowledge limitations or failures is assumed to be a crucial process in failure-based learning approaches (for a review, see Tawfik et al. 2015). The experience of failures is expected to challenge learners' existing mental models, thus preparing them to reflect on and resolve their failures (Tawfik et al. 2015). Hence, it is likely that an awareness of knowledge limitations prepares students for the construction of new knowledge during instruction, as it may help learners to pay attention to the concepts that they need in order to adapt their existing incomplete or erroneous understandings.

Students in a PF setting may already notice their knowledge limitations during the initial problem-solving phase when they struggle or even fail to solve a complex and novel problem canonically (Loibl and Rummel 2014a; Kapur 2016; Newman and DeCaro 2018). The initial problem-solving phase may foster a rather global awareness or feeling of failure in students. However, as PF problems are usually complex and ill-structured, they do not explicitly offer students the opportunity to evaluate their solutions with respect to specific features that are met or not met and to recognize the concrete reasons why their solutions do not work (Loibl et al. 2017). Thus, the initial problem-solving phase does not usually foster students' awareness of their specific knowledge gaps. To achieve this, the delayed instruction phase should build on erroneous student solutions (Loibl and Rummel 2014a; Loibl et al. 2017).

Previous PF studies assessed students' awareness of knowledge limitations by using different self-report measures, such as students' perceived competence (Loibl and Rummel 2015) or their perceived knowledge gaps (e.g. Newman and DeCaro 2018).

Recognition of deep features

The third mechanism (i.e., recognition of deep features) is also expected to emerge during the delayed instruction phase in a PF setting. By contrasting the features of typical erroneous student solutions with each component of the canonical solution during instruction, PF is hypothesized to afford learners the possibility to pay attention to the critical conceptual features of the targeted learning concepts (Loibl et al. 2017; Kapur and Bielaczyc 2012). Moreover, the instructor-led comparing and contrasting activity supports students in

recognizing how these critical features relate to one another and to the canonical solution (Kapur and Bielaczyc 2012). Thereby, the third PF mechanism probably fosters students' deep understanding of the targeted learning concept.

In previous PF studies, evidence for students' recognition of deep features was derived from PF students' performance on posttest items assessing conceptual knowledge. Conceptual knowledge is defined as understanding of the deep features of a domain and of the interrelations between these features and principles in a domain (Rittle-Johnson and Alibali 1999).

In summary, according to the PF literature, for PF to be effective, it appears important that students work on a *complex* problem prior to providing instruction, and that the instruction uses *erroneous student solutions* and *compares and contrasts* the features of student solutions with the components of the canonical solution. This fosters students' activation of prior knowledge, awareness of knowledge limitations, and recognition of deep features (see Table 1). While these three PF mechanisms were derived from research focusing on STEM domains (see the review by Loibl et al. 2017), the features of PF design that are expected to foster the three PF mechanisms are considered to be domain-independent (Kapur 2015). Presumably, therefore, the PF effect and its potential underlying mechanisms should also occur for learning in a respectively designed non-STEM setting. Nevertheless, given the features of the PF design and the mechanisms hypothesized to underlie the PF effect, one could also argue that not every domain is equally appropriate for learning in a PF setting: If the instruction in a PF setting has to compare and contrast the components of the canonical solution with features of erroneous student solutions (Loibl and Rummel 2014a; Loibl et al. 2017; Kapur 2016), then the targeted learning concept needs to have canonical solutions with clearly definable components that can be compared and contrasted with features of erroneous solution ideas. However, this does not imply that the PF problem allows only for one possible correct solution. On the contrary, as mentioned above, the PF problem should be complex and allow students to generate different solution ideas. There should, though, be a canonical or established solution to the problem, the crucial components of which can be presented and explained to students during instruction. Otherwise, PF students would probably have difficulties to detect their specific knowledge gaps and to recognize the deep features of the targeted learning concept during instruction. Therefore, Loibl et al. (2017) hypothesize that the effectiveness of PF may only emerge for learning in *structured* domains, which "allow for clear identification of deep features and evaluation of solution attempts" (p. 712). The following section discusses these domain features in more detail.

Productive failure: domain features

Kapur (2015) and Loibl et al. (2017) emphasize the lack of evidence on whether PF is also effective for learning in domains that are less structured than mathematics and science. The characteristics of such structured domains can be illustrated through the mathematics topic that has been investigated in several previous PF studies, namely standard deviation. In these studies (e.g. Kapur 2014; Loibl and Rummel 2014a), secondary school students received data sets of fictitious athletes and their scores per year. Students were asked to explore the most consistent athlete in scores. During the subsequent instruction, the standard deviation formula was presented as a canonical solution to the previous problem-solving task. The standard deviation formula includes crucial

components that are clearly definable and evaluable, such as taking the square of the distances from each data point to the mean in order to ensure that positive and negative values do not cancel each other out.

In contrast to such STEM-related learning topics, several non-STEM learning topics are likely less structured given the differing nature of knowledge between STEM and non-STEM disciplines. Kagan (2009), for instance, describes that the circumstances under which knowledge can be gained differs between the natural sciences, the social sciences, and the humanities. While the natural sciences can often observe phenomena under controlled conditions, the social sciences and humanities cannot (Kagan 2009). The level of control of the conditions under which knowledge can be gathered within the disciplines may equate to the level of structuredness of the knowledge that is discovered within these disciplines. Specifically, the lower the control of conditions under which certain phenomena can be observed, the higher the possibility to develop multiple and partly contradictory explanations and interpretations of these phenomena, and consequently the greater the difficulty of defining the deep features, rules, and patterns underlying these phenomena and concepts.

The claim that several learning topics in non-STEM domains are not as structured as the above-described STEM learning topic (i.e., standard deviation) may be exemplified through learning topics within the social and educational sciences. The curricula for social and educational science classes in the German state of North Rhine-Westphalia (where the present studies took place), for instance, prescribe learning topics such as social inequality (for social science classes) or learning and development (for educational science classes). In this context, secondary school students should get to know different perspectives, theories, and models, but canonical explanations with clearly evaluable components for the development of social inequality or the processes underlying learning and development cannot be taught (although some teachers may try to do so in classroom realities). Hence, these or similar non-STEM learning topics appear to have little structure and may therefore be inappropriate for learning in a PF setting. However, there are also more structured non-STEM learning topics. For instance, the social and educational science curricula also include interpreting empirical evidence and analyzing experiments by referring to different quality criteria. These learning topics related to social science research methods can be characterized as rather structured, as they allow for a clear identification of crucial principles and criteria relevant for evaluating empirical evidence and experimental study designs, such as the systematic variation of variables, the random assignment of participants, or sample size. Nevertheless, depending on the traditions or culture within a social science discipline, the principles and criteria for an experimental study design appropriate for investigating a certain research question may differ. Thus, one could argue that topics related to social science research methods, such as principles of experimental design, are not as structured as several STEM topics (e.g. standard deviation in mathematics), but are more structured than non-STEM topics relating, for instance, to theoretical, political, or other debates. On a continuum ranging from structured to non-structured, topics related to social science research methods may be positioned towards the structured side of the continuum and may thus be termed as *rather structured* topics. We assume that these or similarly structured non-STEM topics may be appropriate for learning in a PF setting. As previously mentioned, however, research has not focused on the effectiveness of PF for learning in non-STEM domains, which are usually not as highly structured as several STEM-related learning topics. Nevertheless, some scholars have tested the PF effect in domains other than mathematics and science. These studies are described in more detail in the following section.

Design features and domains in previous studies: an overview

Although some studies have tested the effect of problem solving prior to instruction on learning in non-STEM domains, it can be noted that the PF effect has not been tested in what one would call a *classical* PF study in these domains. Classical PF studies, such as those by Kapur (2010, 2011, 2012), Kapur and Bielaczyc (2011, 2012), Kapur and Lee (2009), Loibl and Rummel (2014a, b), and Westermann and Rummel (2012), focus on students' learning in a STEM domain (i.e., mathematics) and additionally share the following three characteristics:

(a) The implemented PF or *problem-solving prior to instruction condition* fulfilled crucial design features described by Kapur and Bielaczyc (2012): (1) students collaboratively invented intuitive solution ideas to a complex problem during the initial problem-solving phase, and (2) erroneous student solutions were compared and contrasted with the canonical solution during the subsequent instruction phase.
(b) The effectiveness of PF was tested for *secondary school students*' (or undergraduate students') knowledge acquisition.
(c) The effectiveness of PF for students' knowledge acquisition was examined by comparing PF to a particular *control condition*; specifically, by comparing PF, a problem solving prior to instruction approach, to an instruction prior to problem-solving approach.

Note that the control condition in classical PF studies is sometimes termed "instruction prior to problem solving" (Loibl and Rummel 2014a, b) or "lecture and practice" (Kapur 2010, 2011; Kapur and Lee 2009), although most studies call this condition "direct instruction (DI)" (Kapur 2012; Kapur and Bielaczyc 2011, 2012; Westermann and Rummel 2012). The main difference between the PF condition and the DI condition is the timing of instruction: While DI students receive instruction *prior* to problem solving, PF students receive instruction *after* problem solving. The DI approach implemented in classical PF studies differs from DI approaches implemented in studies conducted by, for instance, Stevens et al. (1991) or Klahr and Nigam (2004), in which the term *direct* does not relate to the timing of instruction (as in classical PF studies) but rather to the fact that detailed information and explanations are *directly* or *explicitly* provided by the instructor instead of learners having to discover this information on their own. In these studies, direct instruction (in terms of explicit instruction) was given between two learning activities (cf., Klahr and Nigam 2004) or prior to a learning activity, and in addition adaptively during a learning activity (cf., Stevens et al. 1991). In the classical PF studies described here, both DI students *and* PF students receive explicit instructional explanations, either prior to problem solving in the DI condition or after problem solving in the PF condition. Students in both conditions do not usually receive any instructional support *during* the problem-solving activity.

Although most previous PF research focused on the effect of problem solving prior to instruction on learning in STEM domains (for a review, see Sinha and Kapur 2019), eight studies should be mentioned which examined the effect of problem solving prior to instruction on learning in non-STEM domains or on learning domain-general skills (see Table 2). These studies differ from classical PF studies with regard to (a) particular design features of the implemented problem-solving prior to instruction condition, (b) participants' age, or (c) the implemented control condition. Hence, we describe these studies as PF-*similar* studies. The eight PF-similar studies focused on learning in psychology (i.e., schema and

Table 2 PF-similar studies on learning in non-STEM or domain-general settings

	Two studies by Schwartz and Bransford (1998)	Matlen and Klahr (2013)	Glogger-Frey et al. (2015)	Chase and Klahr (2017)	Kant et al. (2017)	Tam (2017)	Marei et al. (2017)
Learning topic	Schema and encoding concepts/ psychology	CVS	Evaluation of learning strategies/ educational psychology	CVS	CVS	Dental hygiene	Dental surgery
Design of problem-solving prior to instruction condition	Analyzing contrasting cases of simplified data from classic psychology experiments prior to a lecture or text	Designing experiments with low guidance (no instructional explanation) prior to high guidance while designing experiments	Evaluating learning strategies in contrasting cases of high school students' learning journals prior to instruction	Interpreting contrasting cases of experimental designs prior to a lecture on the CVS	Conducting a virtual experiment prior to watching a video modeling example	Working on an ill-structured moral dilemma problem (without instructional support) followed by a lecture phase	Interacting with virtual patients (making diagnoses and selecting treatments) followed by a lecture
Collaboration during problem solving	No	No	No	Yes	No	Yes	No
Instruction uses student solutions	No	No	No	No	No	Yes	No
Design of control condition	Reading or summarizing a text about the contrasting cases followed by a lecture	Designing experiments with high guidance followed by design activity with low guidance	Studying the solution in a worked example prior to instruction	Receiving a lecture on the CVS followed by interpreting contrasting cases of experiments	Watching a modeling example followed by conducting a virtual experiment	Receiving a lecture followed by working on a well-structured moral dilemma problem (with instructional support)	Receiving a lecture followed by interacting with virtual patients
Sample group	College students (study 1: $N=21$; study 3: $N=36$)	3rd graders ($N=52$)	Student teachers ($N=42$)	4th and 5th graders ($N=101$)	7th graders ($N=107$)	College students ($N=50$)	College students ($N=84$)

Table 2 (continued)

	Two studies by Schwartz and Bransford (1998)	Matlen and Klahr (2013)	Glogger-Frey et al. (2015)	Chase and Klahr (2017)	Kant et al. (2017)	Tam (2017)	Marei et al. (2017)
Result	Problem solving prior to instruction > control condition	Problem solving prior to instruction = control condition	Problem solving prior to instruction < control condition	Problem solving prior to instruction = control condition	Problem solving prior to instruction < control condition	Problem solving prior to instruction = control condition	Problem solving prior to instruction = control condition

encoding concepts), educational psychology (i.e., learning strategies), medical domains (i.e., dental hygiene and dental surgery), and on learning the Control of Variables Strategy (CVS). The CVS is often framed in a STEM-related context (e.g. in physics by Kant et al. 2017), but can also be described as a domain-general skill of scientific inquiry and reasoning (Chase and Klahr 2017). Of all eight studies, only the two by Schwartz and Bransford (1998) found a positive effect of problem solving prior to instruction on student learning (see Table 2). In the following paragraphs, we briefly describe how the PF-similar studies differ from classical PF studies.

Seven studies (all except Tam 2017) differ from classical PF studies regarding the *design of the problem-solving prior to instruction condition*. In these studies, the instruction phase did not compare and contrast erroneous student solutions with each other and with the canonical solution (as in classical PF studies). As described above, using students' erroneous student solutions and comparing the features of these solutions with the components of the canonical solution during the instruction phase is hypothesized to trigger crucial mechanisms (i.e., awareness of knowledge gaps and recognition of deep features) that may underlie the PF effect (see Table 1). Thus, it remains unclear whether the lack of effects of problem solving prior to instruction shown by the majority of these PF-similar studies is attributable to the differing design of the respective condition compared to classical PF studies. Moreover, in six studies (all except Tam 2017; Chase and Klahr 2017), students worked individually during the initial problem-solving phase and not in small groups or pairs, as in classical PF studies. However, so far, evidence on the role of collaboration for the effectiveness of problem solving prior to instruction, although described as an important feature of the PF design by Kapur and Bielaczyc (2012), is rare and inconclusive (e.g. Weaver et al. 2018).

The studies conducted by Chase and Klahr (2017) and Matlen and Klahr (2013) differ from classical PF studies regarding the *age of the participants*, with participants being younger than in classical PF studies. Other studies testing problem solving prior to instruction compared to instruction prior to problem solving with rather young school students (i.e., 2nd to 5th graders) also found no effect of the former on learning in mathematics (cf., Loehr et al. 2014; Fyfe et al. 2014; Mazziotti et al. 2019). Hence, it may be assumed that the effectiveness of PF does not apply to young students, and that the age of the participants is the reason why Chase and Klahr (2017) and Matlen and Klahr (2013) did not find a positive effect of problem solving prior to instruction on students' learning of the CVS.

The studies by Schwartz and Bransford (1998), Glogger-Frey et al. (2015), and Tam (2017) differ from classical PF studies regarding the *control condition*. Schwartz and Bransford (1998) and Glogger-Frey et al. (2015) compared the problem-solving prior to instruction condition not with an instruction prior to problem-solving condition but rather with another learning activity (e.g. studying worked examples or reading a text) prior to instruction. Thus, regardless of whether students in the problem-solving prior to instruction condition outperformed (Schwartz and Bransford 1998) or did not outperform (Glogger-Frey et al. 2015) their counterparts in the control condition, it remains unclear whether these students would have outperformed students from an instruction prior to problem-solving condition as in classical PF studies. In the study by Tam (2017), students in the control condition worked on well-structured problems after instruction and received instructional guidance during problem solving. Hence, in contrast to classical PF studies, Tam (2017) varied not only the timing of instruction but also the type of problem (ill-structured versus well-structured) and the provision of instructional guidance during problem solving (without versus with guidance). Due to these confounding factors, the study findings are difficult to interpret.

In summary, the eight PF-similar studies can be seen as first attempts to transfer the effect of problem solving prior to instruction from learning in STEM domains to learning in non-STEM domains. As these studies revealed inconsistent findings and differ from classical PF studies, it remains unclear whether the effectiveness of PF depends on the learning domain or rather on particular design features of the problem-solving prior to instruction approach.

Productive failure in learning non-STEM domains

As argued above, the effectiveness of PF is hypothesized to depend on certain design features and learning-related mechanisms (see Table 1). To date, these mechanisms and the effectiveness of PF have only been examined for learning in STEM domains. Only a few studies tested the effect of problem solving prior to instruction on learning in non-STEM domains (see Table 2), and mostly demonstrated no or a negative effect of problem solving prior to instruction on learning in a domain other than mathematics or science. However, these studies differ from the design of classical PF studies, whereby it remains unclear whether the effectiveness of PF depends on the learning domain or rather on certain design features as hypothesized by Kapur (2015). Hence, there is a need to test the effectiveness of PF for learning in a non-STEM domain or—according to Kapur (2015) and Loibl et al. (2017)—in domains that are *less structured* than mathematics and science. At the same time, as mentioned above, it is hypothesized that the PF effect may depend on the structuredness of the learning domain (Loibl et al. 2017). Consequently, probing the effectiveness of PF for learning beyond STEM domains goes along with three challenges and requirements: (1) the design of the PF setting should follow certain principles such that mechanisms hypothesized to underlie the PF effect can emerge, (2) the study design should emulate the features of classical PF studies, and (3) the learning domain should be less structured than mathematics and science, but at the same time meet a certain level of structuredness such that students' solution ideas can be evaluated as well as compared and contrasted with clearly defined features of the canonical solution during instruction. We aimed to face these challenges in the present studies and extended PF from learning in mathematics and science to learning social science research methods. As argued above, topics related to these methods (e.g. analysis of experiments and evaluation of empirical evidence) can be described as rather structured, as they are less structured than the majority of STEM-related learning topics but still allow for a clear definition and evaluation of canonical components. To investigate the PF effect on learning social science research methods, we conducted two quasi-experimental studies while emulating the design of classical PF studies. That is, (a) our *PF condition* followed the features that are expected to trigger the PF mechanisms (see Table 1), (b) we conducted our studies with *secondary school students*, and (c) we compared two conditions: PF with problem solving prior to instruction and *DI* with instruction prior to problem solving.

Research context of the present work

In the state of North Rhine-Westphalia, where the present studies took place, students can usually choose a social or educational science class in 10th grade. According to the German curricula for the respective classes, students should not only acquire content knowledge of

different theories and concepts, but also knowledge related to the research methods and scientific practices within the social and educational sciences. To foster students' knowledge related to research methods and scientific practices, visits to out-of-school labs are expected to be promising (e.g. Pauly 2012). Out-of-school labs were increasingly initiated after the first PISA study in the year 2000, which demonstrated that 15-year-old students' scientific literacy scores in Germany were significantly lower than the overall average of the 43 investigated countries (OECD 2001). To increase students' scientific literacy following this "PISA shock" (Waldow 2009), numerous out-of-school labs for natural sciences and recently also for social sciences were founded in Germany. Out-of-school labs are non-formal learning settings that are assumed to be highly authentic environments for learning scientific ways of thinking and working due to their location (often on a university campus), the instructors (often real scientists or prospective scientists), and the scientifically authentic materials, methods, and contents that students work with during their visit (Garner and Eilks 2015; Scharfenberg and Bogner 2014; Glowinski and Bayrhuber 2011). Usually, projects in such labs run during regular school days; students take part as a whole class and arrive at the lab together with their teachers. As the visits are typically organized by the teachers, they are a compulsive school activity for every student of the class. To enable such formal activity during regular school hours, the lab projects often match the respective curriculum. The present studies took place in an out-of-school lab for social sciences at a large German university.

Investigating the effectiveness of PF for learning social science research methods in an out-of-school lab, as compared to conducting the studies in schools, holds three advantages: (1) As the classes visit the out-of-school lab together with their teachers, the location / environment (lab on a university campus) and the instructor (often a scientist) are the same in all classes. (2) As a non-formal learning setting, the lab may offer a better and more appropriate space for learning by PF than the usual and formal school context, because "school too often allows much less space for risk, exploration, and failure" (Gee 2005a, p. 35). In many cases, experiencing failure has negative connotations in school, and is closely linked to unsatisfactory performance and bad grades (Gee 2005b). (3) Out-of-school labs appear to be ideal settings for implementing PF. More specifically, out-of-school labs aim at implementing authentic experiential learning activities that emulate processes of scientific inquiry in order to situate learners in the role of a scientist and to foster their knowledge about scientific ways of thinking and working (e.g. Euler 2004; Glowinski and Bayrhuber 2011). PF can also situate learners in the role of a scientist by emulating features of authentic scientific practices (Cho et al. 2015; Kapur and Toh 2015). That is, students are asked to explore and to generate solution attempts to a complex problem during an initial problem-solving phase. When being taught the canonical solution in the subsequent instruction phase, students are then required to falsify their initial assumptions about potential solutions. This process of posing hypotheses, discovering that these conjectures are limited or even false, and finally refuting and falsifying the initially generated hypotheses, is characteristic for scientific inquiry processes (cf., Chalmers 2013). Accordingly, it appears that the combination of PF and the out-of-school lab could be particularly beneficial for learning social science research methods.

Research questions

So far, we have argued that problem solving prior to instruction approaches, such as PF, have a greater effect on students' knowledge acquisition than instruction followed by problem solving approaches, such as DI. While this effect has been demonstrated for learning

in STEM domains (especially mathematics), the effectiveness of PF has not been tested for learning in non-STEM domains. Previous research suggests that the PF effect does not depend on the domain (cf., Kapur 2015), but rather on particular features of the PF design (see Table 1), which are expected to foster certain learning-related mechanisms (cf., Loibl et al. (2017) and thus the effectiveness of PF for learning. Therefore, when adhering to these design features, the PF effect should be transferable from learning in STEM domains to learning in non-STEM domains. Building on this argument, the present studies focus on the question of whether the effectiveness of PF unfolds when learning social science research methods. To investigate this question, we conducted two quasi-experimental studies in an interdisciplinary out-of-school lab at a large German university and compared PF to DI.

As the present studies are a first attempt to transfer the design of *classical* PF studies from learning in a STEM domain to learning in a non-STEM domain, we first examine whether our study design fulfilled the features of the classical PF design that are expected to foster the PF mechanisms and thus the effectiveness of PF. Subsequently, we explore whether the mechanisms that are hypothesized to underlie the PF effect, which have also only been replicated for learning in STEM domains, emerge when learning social science research methods. Finally, we investigate our main research question, namely whether the PF effect occurs for learning social science research methods. In summary, we investigate the following two design-related questions, three mechanism-related questions, and one main research question:

- **Design-related question 1:** Does the problem-solving task meet the sweet spot of complexity (cf., Table 1), insofar as PF students are able to invent solution ideas without solving the problem canonically? This design feature is assumed to be important for promoting PF students' activation of relevant prior knowledge during the initial problem-solving phase (cf., Kapur and Bielaczyc 2012; Loibl et al. 2017).
- **Design-related question 2:** Do PF students' solution ideas indeed contain the errors that are focused on during instruction? In the present studies, the instruction lesson was the same in both conditions, and thus standardized by using erroneous student solutions identified in pilot tests as typical for the materials. Using students' erroneous solutions during instruction is expected to foster students' awareness of knowledge limitations and their recognition of the deep features of the targeted learning concept (cf., Loibl and Rummel 2014a; Loibl et al. 2017).
- **Mechanism-related question 1:** Do PF students *activate relevant prior knowledge* during problem solving? According to previous PF studies, this mechanism should be reflected in a positive association between PF students' performance during problem solving (e.g. the quality of their invented solution ideas) and their performance on a posttest assessing their learning outcome (cf., Kapur and Bielaczyc 2012; Loibl and Rummel 2014b). This positive association would indicate that PF students were able to activate relevant prior knowledge during the problem-solving phase, which subsequently helped them to learn the targeted concept.
- **Mechanism-related question 2:** Do PF students develop an *awareness of their knowledge limitations*? Given the findings of Loibl and Rummel (2015), this mechanism should be indicated by PF students' competence perceptions after both learning phases: PF students should report lower perceived competence than DI students after both learning phases. Moreover, PF students' perceived competence after instruction should correlate with their learning outcome. This positive correlation would indicate that the instruction helped students to accurately evaluate their competence.

- **Mechanism-related question 3:** Do PF students *recognize the deep features* of the targeted learning concept during instruction? So far, evidence for this mechanism stems from students' performance on posttest items assessing their deep understanding (cf., Loibl et al. 2017). That is, PF students should perform better than DI students on items testing their ability to apply and transfer their knowledge for solving novel and unfamiliar tasks. This question also addresses our main research question, but focuses on certain items of the posttest assessing deep understanding, while the main research question relates to the total learning outcome.
- **Main research question:** Do students in the PF condition outperform students in the DI condition when learning social science research methods? As ample research has demonstrated the effectiveness of PF for learning in a structured domain, we hypothesize that PF students will outperform DI students on learning social science research methods.

Method study 1

Participants and learning domain

Participants were 10th graders from eight secondary schools in the German state of North Rhine-Westphalia. In total, 213 students from eleven social or educational science classes agreed to participate with written parental consent. 212 students (Age: $M = 16.43$, $SD = 0.78$; 62% girls) who were present during all learning phases were included in the analysis. Our sample is sufficient to reveal effects of $\eta^2 \geq 0.04$ or $d \geq 0.4$ ($f = 0.20$, $1\text{-}\beta = 0.83$; G-Power analysis; Faul et al. 2007).

Participants had no or limited instructional experience with the targeted learning topic prior to the study. The topic for study 1 was principles of experimental design, which is from the 11th grade syllabus. According to the respective curriculum, at the end of the 11th grade, students should be able to analyze experiments with respect to different quality criteria. Thus, we expected that the 10th graders in our study had no instructional experience with the topic of principles of experimental design in the social sciences. Selecting a learning topic from the next year's syllabus is a common strategy employed in previous PF studies to ensure that the study participants have no instructional experience with the given topic (e.g. Kapur 2010).

Experimental design

The eleven participating classes were each randomly assigned to our experimental conditions as a whole. Assignment of whole classes was determined by the out-of-school-lab setting of our study. Our study took place in an interdisciplinary out-of-school lab at a large German university, which offers projects for natural sciences, social sciences, and humanities for students from secondary schools (from 5th grade to 13th grade).

As out-of-school labs are visited by whole classes with their teachers during regular school hours, we used a quasi-experimental design and assigned six whole classes ($n = 121$) to the PF condition and five whole classes ($n = 91$) to the DI condition. Thus, we implemented a between-subjects design. In both conditions, students experienced two successive learning phases: a problem-solving phase and an instruction phase. PF students

experienced the instruction phase *after* the problem-solving phase and DI students experienced the instruction phase *prior* to the problem-solving phase.

Besides the two learning phases, we implemented four test phases, comprising three questionnaires and a posttest. Prior to the first learning phase, a pre-questionnaire assessed students' grades in three different subjects: social sciences, German language, and mathematics. After the first and second learning phase, a questionnaire assessed students' perceived competence. At the end of the study, all students completed a posttest.

Learning materials

Problem-solving tasks and instruction lesson

To teach students principles of experimental design, we created an instruction lesson and two isomorphic problem-solving tasks (Problem 1: a teacher conference scenario, and Problem 2: a parent-teacher conference scenario). Problem 1 was embedded in a cover story about a teacher conference (see Online Resource 1) and depicted three different suggestions and opinions of three fictitious math teachers about how to improve math teaching and learning in 10th grade. Problem 2 was embedded in a cover story about a parent-teacher conference and depicted three different suggestions and opinions of fictitious parents and history teachers about how to improve history education in 9th grade. Both problems asked students to imagine themselves as educational researchers and to design a study which would allow them to investigate all opinions and suggestions mentioned by the fictitious teachers (and parents). The instruction lesson built upon Problem 1 and comprised the following three parts:

- **Part 1:** Recapitulation (in PF) or introduction (in DI) of Problem 1 as depicted in Fig. 1.
- **Part 2:** Comparing and contrasting four features of typical erroneous student solutions with the components of the canonical solution (see Table 3), whereby the canonical solution was explained in a step-by-step procedure. The contents of the second part of the instruction were based on the results of pilot tests of our learning materials, which are described in the section below.
- **Part 3:** Presentation of the canonical solution, namely a 2×2 factorial design with pre- and posttest and questionnaire to measure control variables (see Online Resource 2). The contents of the third part of the instruction phase were based on solutions to Problem 1 developed by experts in our research group. The instructor also presented an alternative canonical solution to Problem 1: a 3×3 factorial design with pre- and posttest in which the control variable (i.e., language skills of students) was implemented as a third factor. However, the instructor described that this would be challenging in real-world learning settings, such as schools, for several reasons. Thus, a 2×2 design with pre- and posttest and questionnaire was presented as being more appropriate.

Pilot tests of the learning materials

In two rounds of pilot tests, we aimed to ensure that the teacher conference problem (Problem 1) had an appropriate level of complexity, allowing students to produce (non-canonical) solution ideas (cf., Kapur and Bielaczyc 2012), and that the instructional explanation used typical erroneous student solutions (cf., Loibl and Rummel 2014a).

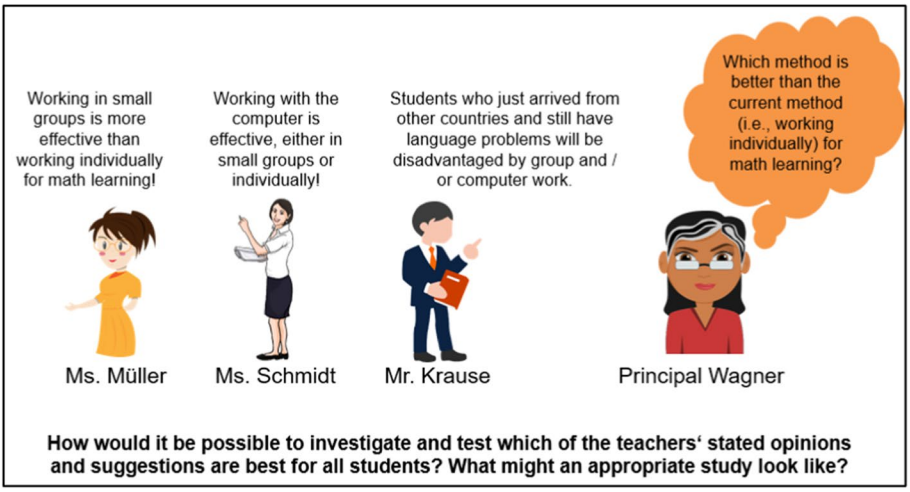

Fig. 1 Recapitulation / introduction of Problem 1 during instruction

Table 3 Features of erroneous solution ideas versus components of canonical solution

Features of typical erroneous student solutions	Components of canonical solution
No inclusion of a control condition: Students often did not plan to compare the suggested learning methods (i.e., collaborative learning and computer-supported learning) to the usual method of learning and practicing individually	*Inclusion of a control condition:* The instructor explained the importance of comparing new methods to the usual method for examining the effectiveness of new methods
No systematic variation of variables: Students often planned to compare groups that varied on more than one factor. For instance, they planned to compare individual computer-supported learning to collaborative learning without computer support	*Systematic variation of the variables:* The instructor explained the relevance of varying only one feature at a time to obtain fairly clear evidence
No administration of a pretest in addition to a posttest: Students often did not think about including a pretest or a certain pre-questionnaire in addition to a posttest or post-questionnaire in their study design	*Administration of a pretest in addition to a posttest:* The instructor explained that the use of a pretest in addition to a posttest is important for investigating students' learning gains and for comparing prior knowledge between the different conditions
No measurement of relevant control variables by using a questionnaire: Students often did not plan to assess whether or not the students had language problems and / or just arrived from other countries	*Measurement of relevant control variables by using a questionnaire:* The instructor explained that the assessment of control variables enables educational scientists to take certain features into account that may have an impact on, for instance, the learning outcome

In the first pilot round, we tested four triads from two 10th grade classes, which were participating in different out-of-school lab projects at the time of our pilot tests. We investigated whether the students were able to generate and explore different solution ideas to Problem 1. According to Kapur and Bielaczyc (2012), the problem-solving task meets a sweet spot of complexity when students are able to generate different solution ideas

without solving the problem canonically. On average, the students generated 6.5 solution ideas in their groups and these solution ideas typically included at least one of the four errors, which are listed in Table 3 (see left column). Building on the findings of this first round of pilot tests, we created the instruction lesson.

In the second pilot round, we tested both learning phases of our PF design with one whole social science class in the out-of-school lab. The results again revealed that 10th graders were able to generate various solution ideas without solving the problem canonically. Specifically, eight student groups developed an average of 4.6 ideas, which again included at least one of the four errors listed in Table 3.

In summary, the results of our pilot tests suggested that the teacher conference problem (Problem 1) met a sweet spot of complexity, as students were able to produce solution ideas without solving the problem canonically. Additionally, these results enabled us to use typical erroneous solution ideas for the instruction lesson.

Measures

To investigate our research questions, we collected and coded students' solutions generated during the problem-solving phase, assessed their perceived competence with a questionnaire after both learning phases, and measured students' learning outcome in a posttest. See Table 4 for an overview of the research questions and the measures used for investigating these questions. Additionally, we administered a pre-questionnaire to assess certain control variables. In the following paragraphs, we describe all measures in more detail.

Pre-questionnaire

Findings from previous PF studies (e.g. Kapur, 2014; Loibl and Rummel 2014b) showed that students' (math) grades influence their learning outcome. Therefore, students' grades were measured as a control variable by asking students to indicate their grades in three different subject areas in the pre-questionnaire administered prior to the first learning phase: social sciences, German language, and mathematics. Students' grades were used as an indicator of their general achievement in the three school subjects. We assumed that students' achievement in these subjects could be related to their ability to work on the problem-solving task and to develop an understanding of the targeted learning concept. That is, the learning topic is strongly related to the *social sciences*, the task that students were asked to solve was a complex *word problem*, and prior knowledge on *mathematical concepts* could be helpful for understanding principles of experimental design, such as the systematic variation of variables.

Student solutions

To analyze whether PF students were able to generate different solution ideas without solving the problem canonically (design-related question 1) and whether the solution ideas contained typical errors that were discussed during instruction (design-related question 2), we coded the solution ideas that students generated during the problem-solving phase, measuring the *quality and quantity of PF student solutions*. This also allows us to investigate whether PF students activated relevant prior knowledge (mechanism-related question 1).

The invented solution ideas were collected after students completed the problem-solving phase. Previously, all groups had received blank sheets of paper for their group

Table 4 Measures used for investigating the research questions in study 1

Questions		Measures
DQ1	Does the problem-solving task meet the sweet spot of complexity?	Student solutions
DQ2	Do PF students' solution ideas indeed contain the errors that are focused on during instruction?	
MQ1	Do PF students activate relevant prior knowledge during problem solving?	
MQ2	Do PF students develop an awareness of their knowledge limitations?	Questionnaires 1 and 2
MQ3	Do PF students recognize the deep features of the targeted learning concept during instruction?	Posttest
RQ	Do students of the PF condition outperform students of the DI condition when learning social science research methods?	Posttest

DQ design-related question, *MQ* mechanism-related question, and *RQ* main research question

work, and students in the PF condition were asked to number their different solution ideas. To measure the quality of the PF solution ideas, we assessed the number of canonical principles within each solution idea and used the best idea (highest number of canonical principles) for further analyses, similarly to previous studies on PF (e.g. Loibl and Rummel, 2014b). The number of canonical principles was composed of two sets of principles (see Table 5), based on both the features of typical student solutions and the components of the canonical solution, which, as mentioned above, had been generated for the teacher conference problem by experts of our research group. One set involves four core principles that—building on typical student misconceptions – were explained in detail during instruction (e.g. inclusion of a control condition). A second set encompasses four further implicit principles of the canonical solution that were not explicitly explained during instruction as they did not represent typical student misconceptions (e.g. comparison of different groups). Therefore, the quality score ranged from 0 (solution without canonical principles) to 8 (canonical solution). See Online Resource 3 for a coding example of a typical PF student solution. Two raters scored around 30% of the student solutions, i.e., $n = 46$ ($ICC_{absolute} = 0.71$; 95% CI [0.48, 0.84]). According to Cicchetti (1994), this ICC value can be interpreted as good.

Questionnaires 1 and 2

To analyze whether PF students became aware of their knowledge limitations (mechanism-related question 2), students' *perceived competence* was assessed after the first and the second learning phase using the short scale of intrinsic motivation developed by Wilde et al. (2009). The questionnaire includes four subscales (interest/enjoyment, perceived competence, perceived choice, and pressure/tension) with a total of 12 items rated from 1 (*strongly disagree*) to 5 (*strongly agree*). To form the perceived competence index, students' ratings on the following three items were averaged: (1) I am satisfied with my performance during this learning phase; (2) I was skilled in the activities during this learning phase; (3) I think I was pretty good at the activities during this learning phase. The internal consistency of the subscale for perceived competence was satisfactory (after learning phase 1: *Cronbach's* $\alpha = 0.84$, after learning phase 2: *Cronbach's* $\alpha = 0.91$).

Table 5 Overview of the codes used for evaluating the quality of student solutions in study 1

Codes used for evaluating the quality of the initial student solutions (cf., Online Resource 3)	
Implicit canonical principles (not focused on during instruction)	Experimental test of learning methods as a type of study
	Comparison of different groups (between-subjects design)
	Correct number and naming of conditions
	Inclusion of a posttest
Explicit canonical principles (typical errors focused on during instruction)	Inclusion of a control group
	Systematic variation of variables
	Inclusion of a pretest
	Consideration of a particular control variable within the planned investigation and analysis

Posttest

To investigate whether PF students recognized the deep features of the targeted learning concept (mechanism-related question 3) and whether they outperformed DI students on acquiring knowledge about principles of experimental design (main research question), a posttest assessed the learning outcome as our dependent variable (i.e., knowledge of the four principles of experimental design conveyed in the instruction phase) after the second learning phase.

Students had 30 min to work on the posttest items, and all completed them on time. Building on Bloom's revised taxonomy (e.g. Krathwohl 2002), four items tested *remembering* and *understanding*, three items tested *application,* and two items tested *analysis* and *evaluation*. The distinction between procedural and conceptual knowledge, as used, for instance, by Loibl and Rummel (2014a) in classical PF studies, builds on a classification developed by Rittle-Johnson and Alibali (1999) and their work on mathematical domains. As we needed a more domain-general classification, we decided to use Bloom's taxonomy to construct our posttest items. Nevertheless, this means that our posttest also allows for testing procedural and conceptual knowledge. As Rittle-Johnson, Siegler, and Alibali (2001) describe, procedural knowledge is often tested by asking learners to solve routine problems that require the acquisition of previously learned solution methods. In contrast, conceptual knowledge is often tested by using novel problems that require learners to invent new solution methods based on their knowledge (Rittle-Johnson et al. 2001). Our posttest also includes both items that ask students to reproduce the concepts and solution methods from the previous instruction phase and items that require students to transfer their knowledge acquired during instruction for solving novel problems.

Two raters scored around 20% of the posttests, i.e., $n = 37$ ($ICC_{\text{absolute}} = 0.99$; 95% CI [0.98, 0.996]). This ICC value can be interpreted as excellent (cf., Cicchetti 1994). After an analysis of the posttest items, we excluded one item from further analyses and created a weighted scale of the remaining eight posttest items with a total score from 0 to 25. A more detailed description of the posttest items and of the item analysis can be found in the supplementary material (see Online Resource 4).

Experimental procedure

On the day of the experiment, students underwent four test phases (i.e., pre-questionnaire, questionnaires 1 and 2, posttest) and two successive learning phases: a problem-solving phase and an instruction phase. While the sequence of these two learning phases differed between the two conditions, the problem-solving tasks, the instruction lesson, and the social surroundings during both phases were the same in both conditions.

PF students' first learning phase was the problem-solving phase, in which they were asked to collaboratively generate different solution ideas to an unfamiliar problem (Problem 1) without any instructional support on relevant or correct problem-solving steps. In the second learning phase, they received instruction: First, the task was recapitulated with the whole class (see Fig. 1). Next, the experimenter compared and contrasted typical erroneous student solutions with the components of the canonical solution (see Table 3). Subsequently, the experimenter gave instruction on the canonical solution and explained relevant components thereof. At the end of the second learning phase, the experimenter

presented a second isomorphic problem (Problem 2) and the whole class generated and discussed its canonical solution with the experimenter.

DI students' first learning phase was the instruction phase, in which the unfamiliar problem (Problem 1) was introduced (see Fig. 1) before they received the same instruction as PF students. That is, the experimenter compared and contrasted typical erroneous student solutions with the canonical solution (see Table 3) and then explained the relevant components of the canonical solution. The second learning phase was the problem-solving phase: Initially, students collaboratively repeated and practiced the canonical problem-solving procedure of the problem that had been presented and discussed during the instruction (Problem 1). Subsequently, the same small student groups practiced the solution procedure on the second isomorphic problem (Problem 2).

Our experimental procedure ensured that the *introductory unfamiliar problem* (i.e., Problem 1, introduced at the beginning of the first learning phase) and the *final familiar problem* (i.e., Problem 2, practiced at the end of the second learning phase) were the same in both conditions. By using typical student solutions during the *instruction phase* – similarly to the procedure of previous PF studies (i.e., Loibl and Rummel 2014a, b)—we were able to apply the same instruction in both conditions. Furthermore, to keep instruction comparable across classes and conditions, the same experimenter gave the instruction in both conditions. The instruction was experimenter-led and included whole-class discussions. Thus, the *social surroundings* during instruction were the same in both conditions. The social surroundings were also the same in both conditions during the problem-solving phase, as students were asked to work in small groups without any instructional support. The small groups were formed by the students themselves, often with seat neighbors. Most of the groups were triads as required by the experimenter, with a few exceptions (two or four group members) due to class size constraints.

Table 6 provides an overview of the experimental procedure including the administration of questionnaires and tests, the breaks, and the respective duration of each learning phase, test phase, and break.

Results study 1

For all analyses, the significance level was set at 0.05. For (co)variance analyses, we used partial η^2 as a measure of effect size. According to Cohen (1988), values < 0.01 represent no effect, between 0.01 and 0.06 a small effect, between 0.06 and 0.14 a medium effect, and values > 0.14 a large effect. For correlation analyses, we used Pearson's *r* as a measure of effect size. According to Cohen (1988), values < 0.10 represent no effect, between 0.10 and 0.30 a small effect, between 0.30 and 0.50 a medium effect, and values > 0.50 a large effect.

Prior analyses

Before investigating all study questions, we ensured that PF and DI students did not differ regarding their reported grades. For this purpose, we conducted a MANOVA with condition as factor and the three grades as dependent variables. There were no differences between the two conditions in the three reported grades (social sciences: $M_{PF} = 2.36$, $SD_{PF} = 0.80$, $M_{DI} = 2.53$, $SD_{DI} = 0.98$, $F(1, 209) = 1.91$, $p = 0.17$, $\eta_p^2 = 0.009$, 90% CI [0.00; 0.04]; German language: $M_{PF} = 2.61$, $SD_{PF} = 0.84$, $M_{DI} = 2.63$, $SD_{DI} = 0.87$, $F(1, 209) = 0.02$, $p = 0.88$,

Table 6 Overview of the experimental procedure in study 1

Learning phase	Test phase	Duration (in min.)	PF condition ($n=121$)	DI condition ($n=91$)
	1	15	Pre-questionnaire	
1		45	Problem solving in small groups (Problem 1)	Instruction in whole class (Problem 1)
	2	10	Questionnaire 1	
		15	Break	
2		45	Instruction in whole class (Problem 1) + practicing Problem 2 in whole class	Problem solving in small groups (Problem 1) + practicing Problem 2 in small groups
	3	10	Questionnaire 2	
		60	Break	
	4	30	Posttest	

$\eta_p^2 = 0.000$, 90% CI [0.000; 0.004]; mathematics: $M_{PF} = 2.83$, $SD_{PF} = 1.04$, $M_{DI} = 2.74$, $SD_{DI} = 1.11$, $F(1, 209) = 0.42$, $p = 0.52$, $\eta_p^2 = 0.002$, 90% CI [0.00; 0.02]).

Design features

To investigate design-related question 1, we analyzed whether PF students were able to invent intuitive solution ideas without solving the problem canonically. This indicates whether the problem-solving task met the sweet spot of complexity and addressed PF students' prior knowledge. The analyses of PF students' solution ideas ($N = 173$) reveal that PF groups ($n = 44$) developed an average of 3.93 ($SD = 1.43$) solution ideas and on average, their best solution ideas involved 5.07 ($SD = 1.29$) canonical principles. None of the groups solved the problem canonically, as their invented solution attempts did not apply eight canonical principles. The highest number of canonical principles was seven, applied by only two PF groups in their best solution idea. Thus, similar to our pilot tests, participants of the PF condition were able to independently generate different solution ideas in small groups without solving the problem canonically. These results suggest that our problem-solving task met an appropriate level of complexity, which is expected to be important for fostering students' activation of prior knowledge during problem solving.

To investigate whether our instruction fulfilled the prerequisite hypothesized to foster students' awareness of knowledge limitations and their recognition of the deep features of the targeted learning concept, we analyzed whether PF students' solution ideas included the typical mistakes that were focused on during instruction (design-related question 2). Descriptive statistics show that the majority of PF students did not include the four particular canonical categories in their solution ideas: A *pretest* was missing in 93% of the generated solution ideas. The principle of a *systematic variation of variables* was violated in 82% of all PF student solutions. The principle *naming and controlling of a particular control variable* was absent in 64% of the invented solutions, and again 64% of the ideas did not include a *control group*. Each of the other four categories (experimental test of learning methods, between-subjects design, posttest, and naming of four learning methods)

was included in 50% or more of the generated solutions. These frequencies of the four most frequently missing concepts within the PF solutions demonstrate that PF students' solution ideas contained the typical mistakes which formed the basis of the developed instructional material and which were focused on during the compare and contrast method (see Table 3 in methods section). Therefore, our PF design fulfilled a relevant prerequisite for fostering students' awareness of knowledge limitations and their recognition of the deep features of the targeted learning concept during instruction.

Activation of prior knowledge and awareness of knowledge limitations

Regarding mechanism-related question 1 (i.e., PF students' activation of prior knowledge), as in previous PF studies (e.g. Loibl and Rummel 2014b), we calculated correlations between solution quality and learning outcome as well as solution quantity and learning outcome for the PF condition. The quality of the PF student solutions (number of canonical principles in best idea) correlated significantly with the individual posttest performance ($n = 121$), $r = 0.33$ (medium effect), $p < 0.001$. The correlation between solution quality and learning outcome on an individual level was calculated by the multiple use of the solution-quality score (i.e., the solution-quality score was used for each group member). As the quality of the solutions was assessed on a group level, we also analyzed the correlation between solution quality and learning outcome on a group level with the mean posttest scores of each group ($n = 44$). The correlation on the group level was also significant, $r = 0.55$ (large effect), $p < 0.001$. Regarding the quantity of invented solutions, we found no significant correlation with the learning outcome, either on the individual ($r = 0.11$, $p = 0.24$) or on the group level ($r = 0.14$, $p = 0.36$). The procedure of analyzing the association between students' learning outcome and the quality or quantity of their generated solutions on both an individual and a group level is in line with methods used in previous PF studies (e.g. Kapur 2012; Loibl and Rummel 2014b). With respect to mechanism-related question 1, these results indicate that the PF students were able to activate relevant prior knowledge, as reflected in the correlation of the quality, and not the quantity, of their solution ideas with their posttest performance.

To investigate mechanism-related question 2 (i.e., PF students' awareness of knowledge limitations), we compared students' perceived competence between the two conditions by calculating a MANCOVA with the factor condition, an averaged index of the three subject grades as covariate (due to correlations with students' perceptions), and both measures of perceived competence as dependent variables. The descriptive statistics for students' perceived competence can be found in Table 7.

For students' perceived competence after the *first learning phase* (assessed after problem solving for PF students and after instruction for DI students), the MANCOVA reveals that students of the PF condition reported significantly higher perceived competence than students of the DI condition, $F(1, 209) = 18.93$ $p < 0.001$, $\eta_p^2 = 0.08$, 90% CI [0.03; 0.15]. For the *second learning phase* (assessed after instruction for PF students and after problem solving for DI students), the MANCOVA reveals that students of the PF condition reported significantly lower perceived competence than students of the DI condition, $F(1, 209) = 62.83$, $p < 0.001$, $\eta_p^2 = 0.23$, 90% CI [0.15; 0.31]. Results of correlation analyses show that the learning outcome of PF students significantly and positively correlated with their perceived competence after the second learning phase ($r = 0.41$, $p < 0.001$), but not after the first learning phase ($r = 0.03$, $p = 0.76$). With respect to mechanism-related question 2, these results indicate that PF students

Table 7 Descriptive statistics for students' perceived competence after the learning phases

Perceived competence	min.–max	PF ($n = 121$)		DI ($n = 91$)	
		M	*SD*	*M*	*SD*
After the first learning phase	1–5	3.53	0.63	3.09	0.86
After the second learning phase		3.02	0.90	3.92	0.82

developed an awareness of their competence limitations during the second learning phase, as they reported substantially less (large effect) perceived competence than DI students. This awareness of limited competence after the second learning phase (and prior to the posttest), was apparently realistic and accurate, as indicated by the positive and medium-sized correlation with PF students' learning outcome. However, PF students did not develop an accurate awareness of their competence limitations during the first learning phase, as they reported higher perceived competence than DI students.

In addition to the previous analyses, we conducted a multiple linear regression analysis to investigate whether PF students' prior knowledge activation (i.e., the quality and quantity of solution ideas) and their awareness of knowledge limitations (i.e., perceived competence after both learning phases) not only correlated with their learning outcome, but also predicted it. The results of the regression analysis, which are presented in Table 8, are in line with the results of our correlation analyses.

The results demonstrate that while PF students' solution quality and their perceived competence after the second learning phase (i.e., after instruction) significantly predicted their learning outcome, the quantity of their solution ideas and their perceived competence after the first learning phase (i.e., after problem solving) did not. With respect to students in the DI condition, the results show that only their perceived competence after the first learning phase (i.e., after instruction) significantly predicted their learning outcome. As DI students developed only one solution (namely the canonical solution) during the problem-solving phase after instruction, the quantity of solution ideas was excluded from the analysis.

Table 8 Summary of regression analyses for variables predicting students' learning outcome in study 1

Variable	PF ($n = 121$)			DI ($n = 89$)		
	B	*SE B*	β	*B*	*SE B*	β
Student solutions						
Quality	1.10	0.34	.27**	1.48	0.90	.18
Quantity	0.23	0.30	.06	/	/	/
Perceived competence						
after the first learning phase	−0.43	0.68	−.05	1.12	0.52	.24*
after the second learning phase	2.16	0.49	.37**	−0.41	0.55	−.09
R^2	.25			.08		
F	9.41**			2.58		

*indicates $p < .05$; **indicates $p < .01$

Deep feature recognition and learning outcome

To investigate mechanism-related question 3 (i.e., PF students' recognition of deep features), we compared students' performance on posttest items testing application and transfer abilities between the two conditions. We calculated an ANCOVA with the factor condition, an averaged index of the three subject grades as covariate (due to correlations with students' performance on the posttest items assessing application and transfer), and a particular set of posttest items as dependent variable. For this purpose, we created a scale of the four tasks (i.e., two unfamiliar application tasks and two analysis / evaluation tasks) that required students to apply the contents of the instruction in novel contexts. The total score of this transfer scale ranged from 0 to 16. The descriptive statistics for students' performance on posttest items assessing application and transfer as well as total learning outcome are presented in Table 9.

The results of the ANCOVA show a medium-sized main effect of condition on the transfer scale, $F(1, 210) = 31.11$, $p < 0.001$, $\eta_p^2 = 0.13$, 90% CI [0.07; 0.20]. Contrary to our expectations, PF students achieved a significantly lower score than DI students (see Table 9). With respect to our mechanism-related question 3, these results indicate that PF students did not recognize the deep features of the targeted learning concept during instruction, as they performed significantly worse on posttest items testing their application and transfer abilities than DI students.

To assess differences in the effect of the experimental condition on students' total learning outcome (main research question), we calculated an ANCOVA with the factor condition and an averaged index of the three subject grades as covariate (due to correlations with students' learning outcome). For this analysis, we used the weighted posttest scale (as described in our method section and Online Resource 4) as dependent variable. The ANCOVA yielded a medium-sized main effect of condition on the total learning outcome, $F(1, 209) = 27.46$, $p < 0.001$, $\eta_p^2 = 0.12$, 90% CI [0.06; 0.19]. Contrary to our hypothesis, PF students did not outperform DI students (see Table 9).

Discussion study 1 and post-hoc analyses

Our results suggest that contrary to our expectations, PF did not have an effect on students' learning of social science research methods, namely on learning principles of experimental design (main research question). Against our hypothesis, DI had a greater impact on

Table 9 Descriptive statistics for students' posttest performance

Posttest performance	min.–max	PF ($n = 121$)		DI ($n = 91$)	
		M	*SD*	*M*	*SD*
On items assessing application and transfer (deep feature recognition)	0–16	5.96	3.19	8.04	2.76
On all items of the weighted posttest scale (total learning outcome)	0–25	10.97	5.22	13.88	3.99

learning principles of experimental design than PF. The question is how this result might be explained. The analysis of the student-generated solutions within the PF condition indicates that the two major design requirements of the learning environment, which are assumed to facilitate the mechanisms underlying the effectiveness of PF (cf., Loibl et al. 2017), were indeed fulfilled: 1. PF students were able to independently invent different solution ideas to a complex problem that addressed their prior knowledge (design-related question 1). 2. Typical mistakes within the student solutions matched the four principles that were explained by comparing and contrasting typical student solutions with the canonical solution during the instruction phase (design-related question 2). Asking students to invent different solution ideas to a complex problem fostered the PF mechanism of activating relevant prior knowledge, which is reflected in the significant correlation between the quality of PF students' solution ideas and their learning outcome (mechanism-related question 1). Additionally, confronting students with incorrect solution ideas during the instruction phase facilitated the PF mechanism of developing awareness of knowledge limitations (mechanism-related question 2). This is especially reflected in the lower perceived competence or higher awareness of competence limitations of PF students compared to DI students after the second learning phase, and additionally in the significant positive correlation between this awareness and PF students' learning outcome (mechanism-related question 2). Nevertheless, PF students seemingly did not recognize the deep features of the targeted learning concept during instruction, as they did not outperform DI students, either on items testing transfer abilities (mechanism-related question 3) or on the total learning outcome (main research question). Given that our study fulfilled all features of the PF design, it is surprising that the PF students did not recognize the deep features of the targeted learning concept or—more generally—that we were unable to replicate the effect of PF on learning social science research methods (more specifically: principles of experimental design).

A possible explanation for our findings may relate to PF students' perceived competence prior to instruction. Against our expectation, PF students' perceived competence after the first learning phase was higher than that of DI students (mechanism-related question 2). Moreover, PF students' perceived competence decreased between the two learning phases (after first learning phase: $M = 3.53$, $SD = 0.63$; after second learning phase: $M = 3.02$, $SD = 0.90$). This is in contrast to the data reported in Loibl and Rummel's (2015) study in the context of mathematics, in which PF students' perceived competence was not only lower than that of DI students after both learning phases, but also *increased* between the two learning phases. It must be noted that the analyses conducted by Loibl and Rummel (2015) did not focus on the development of students' competence perceptions. Thus, our interpretations relate only to their reported descriptive statistics. In our post-hoc analyses, we tested for significance in a repeated measures ANOVA, which yielded a significant and large-sized effect of time of measurement, $F(1, 120) = 23.12$, $p < 0.001$, $\eta_p^2 = 0.21$. As the perceived competence after the first learning phase did not correlate with PF students' learning outcome, either in Loibl and Rummel's study (2015) or in our study, we conducted further post-hoc analyses, exploring whether the decrease in PF students' competence perceptions between the two learning phases was related to their learning outcome. Therefore, we conducted a correlation analysis with PF students' learning outcome and the difference score of their reported perceived competence between the first and the second learning phase. A positive difference score indicates a decrease (after learning phase 1 > after learning phase 2) and a negative score an increase (after learning phase 1 < after learning phase 2). For PF students, the results revealed a negative and medium-sized association (r =—0.36,

$p < 0.001$). Accordingly, the stronger PF students' perceived competence decreased between the two learning phases (i.e., positive difference score), the lower their learning outcome. In summary, the results regarding PF students' perceived competence indicate that PF students did not develop an appropriate awareness of their competence limitations prior to instruction, and may have believed that they had already mastered the task. Consequently, PF students might have experienced a frustrating and overwhelming contrast between their self-assessment prior to instruction (regarding the quality of their solutions and their competence) and their experiences during the following instruction phase, in which they recognized their competence limitations.

To ensure that these findings and especially the lack of effect of PF on learning social science research methods are replicated for topics related to social science research methods other than principles of experimental design, in our second study, we investigated the effectiveness of PF for learning to evaluate causal versus correlational evidence. In the following sections, we describe the method and results of the second study, before discussing the results of both studies in greater detail.

Method study 2

Participants and learning domain

152 students participated with written parental consent in all learning and testing phases. The students were 10th graders (age: $M = 16.11$, $SD = 0.90$; 65% girls) from seven social science classes of six secondary schools in Germany. The sample size is sufficient to reveal small-sized effects of $\eta^2 \geq 0.05$ ($f = 0.23$, $1\text{-}\beta = 0.80$; G-Power analysis).

As in study 1, participants had no or limited instructional experience with the targeted learning topic prior to the study. In study 2, we selected the following learning topic related to social science research methods: evaluating causal versus correlational evidence. According to the curriculum, students should be able to independently interpret empirical data by considering certain quality criteria at the end of 11th grade. Thus, as in study 1 and in previous PF studies (e.g. Kapur 2010), this ensured that the 10th graders in our study had no instructional experience with the learning topic.

Experimental design

Similar to study 1, we implemented a quasi-experimental between-subjects design and randomly assigned the seven classes to the PF condition (three classes: $n = 80$) or the DI condition (four classes: $n = 72$). The assignment of whole classes was determined by the setting of our study: As in our first study, the current study took place in an interdisciplinary out-of-school lab at a large German university.

The design of study 2 was fairly similar to that of study 1: Students underwent two learning phases and four test phases. In contrast to study 1, the second learning phase additionally incorporated a practice phase after instruction in the PF condition and after problem solving in the DI condition. The implementation of a practice phase is in line with the PF design within the classical PF studies by Kapur (2010, 2011, 2012, 2014) and Kapur and Bielaczyc (2011, 2012). See Table 10 for an overview of the experimental design.

Learning materials

To teach students the differences between correlational and causal evidence, we designed two isomorphic problem-solving tasks (i.e., one for the collaborative problem-solving phase and one for the individual practice phase) and an instruction lesson.

Both problems were embedded in a description of a real correlational study within the social and educational sciences. Each problem-solving task was presented to students in a PowerPoint presentation on a laptop and students received the following brief information: (1) the background of the correlational study, (2) the hypothesis of the researchers, which indicated a causal relation between two variables, and (3) the design of the correlational study. On the final slide of the PowerPoint presentation, students were asked to (a) think about all possible results that could be revealed by this kind of a study, (b) think about respective interpretations of that hypothetical results, and (c) discuss and reason whether and why (or why not) one of these results could support the researchers' hypothesis. During the collaborative problem-solving phase, students first individually read the task on a laptop. Subsequently, they were asked to collaboratively write down their solution idea in a joint Word document on a laptop of one of the group members and submit the final document on an online platform. During the individual practice phase, students completed each of these steps (reading the materials, solving the problem, and submitting the solution) individually on laptops. The study that was presented to students during the collaborative problem-solving phase concerned the question of whether the use of media with violent content has an impact on children's aggressive behavior (in the following: violence-in-media task). The problem-solving task in the individual practice phase presented a study investigating whether primary school students' usage of the formal (in German: "Sie") versus informal (in German: "Du") form to address teachers affects their achievement. Usually, the formal "Sie" is used to address strangers, people in authority, superiors, etc. and the informal "Du" is used when speaking to friends, family members, and people who offered the "Du".

As in the first study, the instruction lesson used typical student solutions and compared and contrasted features of these solutions with the canonical solution. For this purpose, we again conducted pilot tests of our learning materials, which revealed that 10th graders often developed the assumption that the correlational study presented in the violence-in-media task could reveal causal evidence. Thus, the instruction focused on explaining the differences between causal and correlational evidence. In contrast to our first study, the instruction lesson also introduced failure and mistakes as typical features of scientific inquiry.

Table 10 Overview of the experimental design in study 2

Learning phase	Test phase	PF condition ($n = 80$)	DI condition ($n = 72$)
	1	Pre-questionnaire	
1		Problem solving	Instruction
	2	Questionnaire 1	
2		Instruction (+ practice)	Problem solving (+ practice)
	3	Questionnaire 2	
	4	Posttest	

Thereby, we aimed at reducing students' potential frustration when confronted with their erroneous solution ideas and their own previous failure during instruction. The structure of the instruction lesson was as follows:

(1) Introduction of the steps of a typical scientific inquiry process within the social sciences, and emphasis on the notion that failure is an inherent and constructive part in each of these steps (also within the process of interpreting empirical data);
(2) Recapitulation of the violence-in-media task in the PF condition or introduction of the background of the violence-in-media task in the DI condition;
(3) Presentation of three possible results that could be demonstrated by the violence-in-media study and that were typically named by students in our pilot tests, that is, no association, a negative association, and a positive association between the usage of media with violent content and children's aggressive behavior;
(4) Description that students typically interpret an association between these two variables as causal evidence, and explanation of the issues related to such a misinterpretation (e.g. presence of further confounding factors that may underlie the association of two variables);
(5) Description that the scholars who conducted the violence-in-media study also misinterpreted their correlational results as causal evidence and that such misinterpretations of correlational evidence are widespread in media reports on empirical findings;
(6) Presentation of a canonical and ideal study design (i.e., controlled experiment) for investigating whether the usage of media with violent content has an impact on children's aggressive behavior;
(7) Summary.

Measures

As in study 1, we collected and coded students' solutions generated during the problem-solving phase in order to investigate our two design-related questions and mechanism-related question 1. To examine mechanism-related question 2, we assessed students' perceived competence with a questionnaire after both learning phases. Students' learning outcome was measured in a posttest in order to investigate mechanism-related question 3 and our main research question. A pre-questionnaire again assessed students' grades in three subjects (i.e., social sciences, German Language, and mathematics). Moreover, due to the implemented practice phase, we coded students' solutions generated during the practice phase. As the questionnaires (i.e., pre-questionnaire and questionnaires 1 and 2) were the same as in study 1, we only describe the posttest and the coding of the student solutions in the following paragraphs.

Student solutions

We coded the quality of students' solutions to the violence-in-media problem and of the solutions they had generated to an isomorphic problem during the practice phase. For both coding procedures, we used the same rating scheme, which is illustrated in Table 11.

The codes in the rating scheme are based on the components of the canonical solution that were explained during the instruction phase. The instruction focused on the following contents: (a) by conducting a correlational study, three results can be revealed (i.e., a positive, a negative, or no correlation between two variables), (b) these results cannot be

Table 11 Coding scheme for assessing the quality of student solutions in study 2

Task instruction	Code description	Scores (min.–max.)
Think about possible results that could be revealed from this kind of study	Participants somehow describe the following three results: positive, negative, and no association/relation/connection between the two investigated variable	One point per correctly named result (0–3)
Think about respective interpretations of the hypothetical results	Participants correctly interpret the hypothetical results. That is, they describe no causal relation between the two variables by not using terms such as, because, influence, impact, through, by, etc	One point per correctly interpreted result (0–3)
Discuss and reason whether and why (or why not) one of these results could support the researchers' hypothesis	Participants describe that none of the hypothetical results could support the researchers' assumption about a causal relation between two variables. Participants further give one of the following explanations: (a) correlational studies do not allow for causal conclusions or (b) further factors could underlie the association	One point for "none of the results" and one point for giving a correct explanation: (0–2)

interpreted as causal evidence, and (c) a causal relation between two variables cannot be tested by conducting a correlational study due to confounding factors that could underlie the correlation. As depicted in Table 11, the total score of the solution quality ranged from 0 to 8. Two raters coded 20% of the solutions that students had generated during the problem-solving phase, i.e., $n = 12$ ($ICC_{absolute} = 0.83$; 95%-CI [0.42, 0.95]), as well as 20% of the solutions that students had generated during the practice phase, i.e., $n = 30$ ($ICC_{absolute} = 0.78$; 95% CI [0.55, 0.90]). The interrater reliability of both ratings can be interpreted on the basis of the ICC values and according to Cicchetti (1994) as satisfactory.

Posttest

To investigate our mechanism-related question 3 and our main research question, a posttest assessed students' learning outcome as our dependent variable. The posttest was administered after the second learning phase and a one-hour lunch break in both conditions. The test consisted of six task sets with a total of ten items that students had to individually solve in a prescribed order on a laptop. The tasks assessed students' ability to reproduce, apply, and transfer the contents that were focused on during instruction. The majority of the items asked students to demonstrate their understanding of the targeted learning concepts by answering open questions in their own words. Two raters coded around 20% of the posttests, i.e., $n = 35$. The interrater reliability can be interpreted on the basis of the ICC value ($ICC_{\text{absolute}} = 0.93$; 95%-CI [0.85, 0.96]) and according to Cicchetti (1994) as satisfactory. After an analysis of the posttest items, we excluded two tasks from further analyses and created a joint scale of the remaining eight posttest items with a total score from 0 to 21. A more detailed description of the posttest items and of the item analysis can be found in the supplementary material (see Online Resource 5).

Experimental procedure

As in study 1, students in both conditions participated in two learning phases, and completed three questionnaires (i.e., prior to learning phase 1, after learning phase 1, and after learning phase 2) and a posttest at the end of the experiment.

The *problem-solving phase* lasted for 35 min. It was the first learning phase in the PF condition and took place in the second learning phase in the DI condition. In both conditions, the same problem was presented: Students were asked to collaboratively think about findings and conclusions that could be drawn from a correlational study on the association between children's use of violent media and their aggressive behavior. Students in both conditions were asked to form small groups (i.e., triads). Due to class size constraints, some groups consisted of two to four members rather than three. In both conditions, students did not receive any instructional support on relevant or correct problem-solving steps. Numerous pilot tests ensured that PF students were able to generate solution ideas without solving the problem canonically prior to instruction, which is suggested to be a crucial principle of the PF design (cf., Kapur and Bielaczyc 2012). It should be noted that DI students, who underwent the problem-solving phase after instruction, only had to apply the instructed problem-solving steps in order to solve the problem canonically (which is similar to the procedure in study 1 and previous PF studies). Thus, the problem-solving phase in the DI condition is rather a practice phase, especially as DI students were asked to solve the same problem that was focused on during the previous instructional explanation.

However, by asking PF *and* DI students to work on the same task, we kept the problem-solving phase comparable across both conditions.

By conducting and iteratively adapting the problem-solving task in a series of pilot tests, we were able to use typical student solutions during the *instruction phase*. Students typically developed one, two, or all of the following three hypothetical results during the problem-solving phase: a positive association between students' use of media with violent content and their aggressive behavior, a negative association, and no association. They further interpreted a positive association between the two variables as causal evidence. Therefore, the same experimenter in both conditions focused on appropriate interpretations of the three hypothetical results and on the issue of causality and correlation during the instructional explanation. The instructor also emphasized the commonness of this mistake in media and research, and introduced mistakes and failures as a constructive and normal part of scientific inquiry processes. By using typical student solutions, we were able to apply the same instruction in both conditions. The instruction lasted for 25 min and was the first learning phase in the DI condition and took place in the second learning phase in the PF condition.

After completing the problem-solving and the instruction phase, students in both conditions worked individually on a problem (isomorphic to the problem presented before) during a *practice phase* consisting of two parts: 30 min problem solving followed by 10 min instruction, in which the canonical solution of that second problem was briefly presented by the experimenter.

Results study 2

Prior analyses

Before investigating our research questions, we ensured that PF and DI students did not differ in their prior knowledge (indicated by their grades). For this purpose, we conducted a MANOVA with the factor condition and students' grades in three subjects as dependent variables. There were no differences between the two conditions in the three reported grades (social sciences: $M_{PF}=2.72$, $SD_{PF}=0.86$, $M_{DI}=2.50$, $SD_{DI}=0.81$, $F(1, 148)=1.89$, $p=0.10$, $\eta_p^2=0.02$, 90% CI [0.00; 0.06]; German language: $M_{PF}=2.82$, $SD_{PF}=1.14$, $M_{DI}=2.71$, $SD_{DI}=1.01$, $F(1, 148)=0.46$, $p=0.53$, $\eta_p^2=0.003$, 90% CI [0.00; 0.03]; mathematics: $M_{PF}=2.44$, $SD_{PF}=0.97$, $M_{DI}=2.23$, $SD_{DI}=0.85$, $F(1, 148)=2.04$, $p=0.12$, $\eta_p^2=0.02$, 90% CI [0.00; 0.06]).

Design features

To investigate whether PF students were able to generate a solution idea to the violence-in-media problem without solving it canonically (design-related question 1), and whether they made the mistake that was focused on during instruction (design-related question 2), we analyzed PF students' solution ideas that they had collaboratively generated during problem solving.

The analysis of student solutions shows that 96% of the PF groups ($N=28$) developed the hypothetical result of a positive association between the two variables (i.e., usage of media with violent content and showing aggressive behavior); 50% of the groups thought of a negative association between the two variables, and 64% of the groups stated that the

study could also reveal that the two variables are not linked. However, 93% of the groups stated that the correlational study could reveal causal evidence. Hence, PF students were able to generate intuitive solution ideas without solving the problem canonically (design-related question 1). Moreover, the majority of the PF students generated the three hypothetical results that were discussed during instruction and most of the PF students made the mistake (i.e., misinterpretation of correlational evidence as causal) that was focused on during the instructional explanation (design-related question 2). Therefore, our second study fulfilled the design features that are expected to foster the three learning-related mechanisms. In the following sections, we investigate whether these mechanisms occurred in study 2.

Activation of prior knowledge and awareness of knowledge limitations

To investigate mechanism-related question 1 (i.e., PF students' activation of prior knowledge), we calculated correlations between the quality of students' collaboratively generated solution idea during problem solving and (a) the quality of their individually generated solution idea after instruction (i.e., during the practice phase) and (b) their learning outcome. The results reveal for (a) a significant and small-sized correlation between the quality of PF students' collaboratively generated solution ideas prior to instruction and their individually generated solution after instruction ($r = 0.29$, $p = 0.01$). The correlation analysis revealed for (b) a significant and small-sized correlation between the quality of PF students' collaboratively generated solution idea and their learning outcome ($r = 0.26$, $p = 0.02$). See Table 12 for the descriptive statistics of the quality of the solutions that students generated during the problem-solving and practice phase. It should be noted that the correlations between collaborative and individual performance (during the practice phase and on the posttest) were calculated on an individual level by the multiple use of the collaborative performance score (i.e., the score of the collaborative performance was used for each group member).

To investigate mechanism-related question 2 (i.e., PF students' awareness of knowledge limitations), we conducted a MANOVA with the factor condition and both measures of perceived competence as dependent variables. Table 13 shows the descriptive statistics for students' perceived competence.

For students' perceived competence *after the first learning phase*, the MANOVA reveals a significant and small-sized effect of condition, ($F(1, 150) = 5.33$, $p = 0.02$, $\eta_p^2 = 0.03$, 90% CI [0.00; 0.09]). PF students reported significantly higher competence perceptions than DI

Table 12 Descriptive statistics for students' task performance during the learning phases

	min.–max.	PF (n = 80/79)[a]		DI (n = 72/71)[a]	
		M	*SD*	*M*	*SD*
Collaborative task performance (during the problem-solving phase)	0–8	3.18	1.10	4.57	1.64
Individual task performance (during the practice phase)		4.25	2.24	5.08	1.76

[a]One missing value with respect to the individual task performance in each condition

Table 13 Descriptive statistics for students' perceived competence after the learning phases in study 2

Perceived competence	min.–max	PF ($n=80$)		DI ($n=72$)	
		M	*SD*	*M*	*SD*
After the first learning phase	1–5	3.55	0.74	3.25	0.86
After the second learning phase		3.41	0.72	3.48	0.76

Table 14 Summary of regression analyses for variables predicting students' learning outcome in study 2

Variable	PF ($n=80$)			DI ($n=72$)		
	B	*SE B*	β	*B*	*SE B*	β
Solution quality (problem-solving phase)	0.76	0.35	.23*	0.29	0.28	.12
Perceived competence						
After the first learning phase	−0.74	0.51	−.15	0.71	0.58	.16
After the second learning phase	1.52	0.53	.30**	0.21	0.66	.04
R^2	.17			.05		
F	5.14*			1.16		

*Indicates $p<.05$; **indicates $p<.01$

students after the first learning phase (see Table 13). For students' perceived competence *after the second learning phase*, the results show no effect of condition, ($F(1, 150)=0.29$, $p=0.59$, $\eta_p^2=0.002$, 90% CI [0.00; 0.03]). PF students' perceived competence did not significantly differ from that of DI students after the second learning phase (see Table 13). Correlation analyses demonstrated that PF students' competence perceptions after the first learning phase did not correlate with their learning outcome ($r=-0.11$, $p=0.32$), but their perceptions after the second learning phase did correlate with their learning outcome ($r=-0.31$, $p=0.006$).

The previous results of the correlation analyses are in line with the results of a multiple linear regression analysis, summarized in Table 14.

Deep feature recognition and learning outcome

To investigate mechanism-related question 3 (i.e., PF students' recognition of deep features), we compared students' performance on posttest items testing application and transfer abilities between the two conditions. We conducted an ANCOVA with the factor condition, an averaged index of the two subject grades as covariate (due to correlations with the transfer scale), and a particular set of posttest items as dependent variable. For this purpose, we created a scale of the three tasks that required students to apply the contents of the instruction in novel contexts. The total score of this transfer scale ranged from 0 to 13. Table 15 shows the descriptive statistics for students' posttest performance.

The results of the ANCOVA reveal that PF students did not differ significantly from DI students on the transfer scale ($F(1,149)=0.09$, $p=0.77$, $\eta_p^2=0.001$, 90% CI [0.00; 0.02]).

To investigate our main research question (i.e., do PF students outperform DI students on the learning outcome?), we conducted an ANCOVA with the factor condition, students' total learning outcome (i.e., posttest scale as described in the Online Resource 5) as dependent variable, and an averaged index of two grades as covariate. Students' grades in two subjects (i.e., social sciences and German language) significantly correlated with their learning outcome and were therefore included as covariate. The ANCOVA revealed no effect of condition on students' learning outcome ($F(1, 148) = 0.32$, $p = 0.57$, $\eta_p^2 = 0.002$, 90% CI [0.00; 0.03]). That is, PF students did not outperform DI students on the total learning outcome (see Table 15). In light of these results, our hypothesis of a PF effect has to be rejected.

Discussion study 2 and post-hoc analyses

The results of study 2 replicate the results of study 1 as our findings showed no effect of PF on students' learning about social science research methods. Nevertheless, in contrast to study 1, PF students did not achieve a significantly lower score on the posttest than DI students. The analyses of PF students' solutions generated during the problem-solving phase (design-related questions 1–2) showed that our current study fulfilled major *requirements of the PF design*: PF students were able to generate solution ideas without solving the problem canonically during the initial problem-solving phase (design-related question 1). The subsequent instruction used typical student solutions and addressed typical mistakes (design-related question 2). These design features are expected to foster students' activation of relevant prior knowledge, their awareness of knowledge limitations, and their recognition of the deep features of the targeted learning concept. Students' *activation of prior knowledge* (mechanism-related question 1) was also evidenced by the positive correlation between the quality of PF students' collaboratively generated solution idea prior to instruction and their individually generated solution idea after instruction as well as their learning outcome. These findings indicate that, as in study 1, the initial problem-solving phase in the current study helped PF students to activate relevant prior knowledge for subsequent learning during instruction. With regard to students' *awareness of knowledge limitations* (mechanism-related question 2), our analyses of students' perceived competence show that PF students did not become aware of their competence limitations after the first learning phase and that this lack of perceived competence limitations did not correlate with PF students' learning outcome. Both findings are in line with the results of our first study.

Table 15 Descriptive statistics for students' posttest performance

Posttest performance	min.–max	PF ($n = 80$)		DI ($n = 72$)	
		M	*SD*	*M*	*SD*
On items assessing application and transfer (deep feature recognition)	0–13	5.33	2.35	5.49	2.63
On all items of the weighted posttest scale (total learning outcome)	0–21	9.61	3.62	10.07	3.87

In contrast to the results of our first study, PF students did not report lower perceived competence than DI students after the second learning phase. Nevertheless, PF students' perceived competence after the second learning phase correlated positively with their learning outcome, which again corresponds to the results of study 1. Thus, this finding cannot be interpreted as evidence for a lack of PF students' awareness of their competence limitations after the second learning phase, as their competence perceptions were probably appropriate, as indicated by the positive correlation with their learning outcome. Hence, as in study 1, our second study fulfilled the PF design features that are expected to trigger the PF mechanisms. Moreover, PF students activated relevant prior knowledge during the first learning phase and developed accurate competence perceptions during the second learning phase. Apparently, however, they did not recognize the deep features of the targeted learning concept (mechanism-related question 3), and more generally, did not outperform DI students on learning to evaluate correlational versus causal evidence (main research question).

A possible explanation for this lack of PF effect may again be related to PF students' lack of perceived competence limitations prior to instruction (mechanism-related question 2), which probably accounted for why PF students perceived competence again did not increase between the two learning phases but remained relatively stable. A repeated measures ANOVA revealed a small and non-significant effect of time of measurement, $F(1, 79) = 1.59$, $p = 0.21$, $\eta_p^2 = 0.02$. Thus, PF students' perceived competence after the first learning phase ($M = 3.55$, $SD = 0.74$) did not significantly differ from their perceived competence after the second learning phase ($M = 3.41$, $SD = 0.72$). A post-hoc correlation analysis replicates the finding of study 1, namely a positive correlation between the development of PF students' perceived competence and their learning outcome ($r = -0.31$, $p = 0.005$). In the following section, we discuss the results of our two studies in more detail.

General discussion

Ample research has demonstrated the effectiveness of PF (or problem solving prior to instruction more generally) for learning in different STEM domains (especially mathematics). In a PF setting, students independently invent solutions to a complex and novel problem before receiving instruction that compares typical erroneous student solutions with the canonical solution (Kapur and Bielaczyc 2012). These major features of the PF design are expected to foster three mechanisms that are hypothesized to underlie the PF effect (Loibl et al. 2017): activation of prior knowledge, awareness of knowledge limitations, and recognition of deep features (see Table 1). Although it is claimed that these features are domain-independent (Kapur 2015), so far, the PF effect has not been investigated on learning in non-STEM domains. Against this background, the present work examined the effectiveness of PF for learning in a non-STEM domain. In two studies, we compared the effects of PF and DI on learning social science research methods. We first examined whether our study design fulfilled major features of the PF design (design-related questions 1–2) and whether the mechanisms hypothesized to underlie the PF effect occurred in both studies (mechanism-related questions 1–3). Subsequently, we investigated whether PF students outperformed their counterparts in the DI condition on a posttest (main research question).

As previously summarized, our findings showed that both studies fulfilled major features of the PF design (design-related questions 1–2). Moreover, in both studies, but especially in study 1, crucial PF mechanisms (mechanism-related questions 1–2) occurred,

namely activation of prior knowledge (during problem solving) and awareness of knowledge limitations (after instruction). Nevertheless, PF students did not outperform DI students either on items testing their deep understanding (mechanism-related question 3) or on the total learning outcome (main research question). In the discussions above, we hypothesized that the lack of effect of PF on learning social science research methods found in our two studies may be linked to PF students' perceived competence prior to instruction and particularly to the development of their perceived competence between the two learning phases. Specifically, in contrast to the findings of Loibl and Rummel (2015), PF students' perceived competence was higher than that of DI students after the first learning phase (mechanism-related question 2). Moreover, PF students' perceived competence decreased between the two learning phases in study 1 and was fairly stable in study 2. While students' perceived competence prior to instruction did not correlate with their learning outcome (mechanism-related question 2), the development of their perceived competence between the two learning phases did. The decrease in PF students' perceived competence in study 1 and the stability in study 2 is contrary to the fundamental idea of PF, namely the preparation for future learning (cf., Schwartz and Martin 2004). That is, learners may initially struggle and fail, but subsequently benefit from these experiences (Kapur 2008). The lack of increase in perceived competence may be mirrored in the lack of preparation-for-future-learning effect in our two studies, as PF students did not perceive higher competence after instruction than before. As PF students reported rather high perceived competence prior to instruction, they seemingly did not experience struggle and failure during problem solving, although they actually failed to solve the problem canonically. In the following sections, we discuss possible explanations for why PF students in our two studies did not perceive appropriate competence limitations prior to instruction and why the PF effect did not emerge.

Lack of feedback, feelings of ownership, and lack of actual competence

One explanation for why students in our PF condition did not perceive appropriate competence limitations prior to instruction may lie in the problem-solving task, which differs from previous PF studies. In the studies by Kapur (e.g. 2014) or Loibl and Rummel (e.g. 2014b), 9th or 10th graders received data sets of fictitious soccer or basketball players and their scored points per year. Students were asked to explore the most consistent athlete in scoring points through various mathematical strategies. The data sets were designed such that when students used incorrect procedures to calculate athletes' consistency (e.g. mean or range instead of standard deviation), they would get the same result for each athlete. Therefore, students in these studies might have been more likely to experience failures, knowledge gaps and limitations, and dissatisfaction with and doubts about their solution ideas than in the present studies. It may therefore be important for the PF effect that students receive feedback through their generated solution ideas (cf., Sinha and Kapur 2019). A similar conclusion was drawn by Matlen and Klahr (2013) and Kant et al. (2017), who demonstrated a non-effect of problem solving prior to instruction on students' learning of the CVS for scientific experimentation (see Table 2). Both concluded that, in contrast to classical PF studies, their materials did not enable students to develop an awareness of failure prior to instruction, as students did not receive feedback through their generated solution ideas. Excerpts of student discussions in a study by Kapur (2010), in which 7th graders tried to solve a mathematical problem about rate and speed, demonstrate that PF students did become aware of the disadvantages of their solution ideas during the problem-solving

phase and that this awareness led to the exploration of other solution ideas. For instance, one group member stated "We tried so many methods but in the end, all the methods could not be used" (see Kapur 2010, p. 538). Such a problem-solving process, which requires learners to try out whether their generated solutions work and to adjust their failed solution attempts, features well-structured problems (Jonassen 1997). However, according to Kapur (2008) and Kapur and Bielaczyc (2012), the PF design should ask students to explore solution ideas to a complex and rather ill-structured problem. When solving ill-structured and thus rather open-ended problems, learners cannot usually try out and then adjust their solutions, but are rather asked to justify their solution ideas by referring to their own arguments and personal beliefs (Jonassen 1997). As the present two studies emulated the PF design features described by Kapur and Bielaczyc (2012), the implemented problem-solving tasks in our PF conditions are rather ill-structured problems. That is, PF students were unable to try out and then adjust their solution ideas, and instead had to discuss, justify, and elaborate on their solution ideas within small groups. However, if the PF effect depends on the design of the problem-solving task that promotes PF students' awareness of failure during the problem-solving phase (which is indicated by our two studies but has not been stated in previous literature on PF), then—in contrast to the PF design features described by Kapur and Bielaczyc (2012)—the PF problem-solving task probably needs to be well-structured to enable PF students to receive feedback trough their generated solution ideas.

As PF students in our two studies did not become aware of their failure during the problem-solving phase, it might be assumed that the so-called IKEA effect (Norton et al. 2012) was at play. That is, people tend to "overvalue their (often poorly constructed) creations" (Norton et al. 2012, p. 453) due to feelings of ownership and the invested effort and labor. As such, PF students might have developed feelings of ownership of their solution ideas that they had independently generated prior to instruction, thus overvaluing the quality of their solutions and of their competence during problem solving. Consequently, PF students might have been unprepared to accept the canonical solution during instruction. Glogger-Frey et al. (2015) suggested a similar explanation for their findings demonstrating a negative effect of problem solving prior to instruction (see Table 2). The assumption that PF students might have clung to their collaboratively invented solution ideas during and after instruction can partly be supported by our data: In both studies, PF students were rather unable to canonically solve a familiar posttest task (i.e., a task isomorphic to the task of the problem-solving phase) after instruction. Only 24% of PF students in study 1 and 9% of PF students in study 2 were able to canonically solve a familiar posttest task, while 48% of DI students in study 1 and 18% of DI students in study 2 were able to do so. However, it should be noted that our coding both of students' collaboratively generated solution ideas during problem solving and their individual posttest performance aimed at analyzing the number of canonical components that were included in students' solutions. Therefore, the aforementioned results only suggest that PF students' solutions prior to instruction (i.e., problem-solving phase) and their solutions after instruction (i.e., posttest phase) included similar mistakes.

Another explanation for why PF students in our two studies did not become aware of their competence limitations prior to instruction relates to the Dunning-Kruger effect. This effect describes the phenomenon of learners who are unskilled but unaware of their competence limitations and even overestimate their competence (Kruger & Dunning 1999). Kruger and Dunning (1999) demonstrated that the ability to become aware of one's limited competence increases with the level of actual competence that learners acquire. Our results support this assumption, as PF students' competence perceptions were more appropriate after instruction, i.e., after actually learning the canonical

solution, than before. That is, in both studies, PF students' perceptions after instruction correlated positively with their learning outcome, while their perceptions after the problem-solving phase did not. This positive correlation indicates that students who reported higher perceived competence after instruction indeed performed better on the posttest than students who reported lower perceived competence.

While these potential explanations for our findings either refer to the design of our two studies (see explanation regarding the lack of feedback) or to empirical evidence provided by our studies (see explanations regarding feelings of ownership and lack of actual competence), the explanations suggested below represent rather theoretical conjectures.

Cue utilization and facilitated problem solving

Given the lack of feedback through one's generated solution ideas, the feelings of ownership of self-generated solution ideas, and/or the lack of actual competence, it may be conjectured that PF students were unable to use predictive cues in order to make accurate judgments of their competence during the initial problem-solving phase. According to the cue utilization framework (Koriat 1997), the accuracy of learners' monitoring judgments about their performance or understanding prior, during, or after learning activities depends on their use of predictive cues (e.g. De Bruin et al. 2017). Often, students use non-predictive cues for evaluating the quality of their performance, which is usually accompanied by overconfidence (e.g. De Bruin and van Merrienboer 2017). Non-diagnostic cues, such as speed of generating solutions or accessibility (i.e., ability to recall some kind of information regardless of the quality of the recalled information), do not predict the actual quality of students' learning (cf., De Bruin et al. 2017). PF students in our study might also have relied on the cues of accessibility and speed for their competence judgments, meaning that they were able to generate some kind of solution idea and developed their ideas in the prescribed time during the initial problem-solving phase. To help students to accurately self-evaluate the quality of their performance, or in other words to foster their use of diagnostic cues, it is assumed that feedback, prompts, or some kind of support could be beneficial (De Bruin et al. 2017).

However, facilitating PF students' problem-solving process more generally has been demonstrated to be equally effective (Loibl and Rummel 2014b) or even less effective (Kapur 2011) than withholding any instructional support during the initial problem-solving phase for PF students' learning in mathematics. Nevertheless, when learning in other domains (or with other problem-solving tasks), it may be necessary to facilitate students' problem-solving process in order to trigger their evaluation of their ideas and awareness of their knowledge gaps. Analyses regarding the facilitation of medical students' problem-based learning (Hmelo-Silver and Barrows, 2006), for instance, showed how certain strategies used by the facilitator initiated students' evaluation and in turn the further development of their solution ideas during the problem-solving process.

Hence, future studies should investigate whether facilitating PF students' problem-solving process and / or fostering their use of predictive cues promotes their awareness of failure during problem solving and the effectiveness of PF for learning social science research methods (or in non-STEM domains in general).

Epistemological beliefs

PF students' use of non-predictive cues could also be due to their epistemological beliefs about social sciences. As Pauly (2012) claims, students often assume that social scientists merely express their individual opinions and thoughts, instead of systematically discovering findings in the same way as natural scientists. Thus, the PF students in our studies, who were asked to imagine themselves as social or educational scientists during the problem-solving phase, might have seen no need to critically and systematically evaluate their solution ideas. Instead, they might have perceived themselves as fairly competent, believing themselves to be expressing their individual opinions and thoughts in the same way as social scientists. Due to inappropriate epistemological beliefs, PF students might also have not perceived the problem-solving activity as an authentic simulation of scientific inquiry processes, which require careful and critical evaluation of the executed steps. We investigated students' perceived authenticity using data from the first study, and our findings indeed revealed that PF students did not report higher perceived authenticity than those in the DI condition (cf., Nachtigall et al. 2018).

In addition to potentially less sophisticated epistemological beliefs about social sciences, our participants might have had misconceptions relating to the goals of empirical investigations which they were asked to design in study 1 and to evaluate in study 2. As Chase and Klahr (2017) describe, students often believe that experimental studies aim at guaranteeing a desired outcome and not at examining the impact of specific factors on an outcome. The benefits of PF (or problem solving prior to instruction more generally) approaches, however, may rather relate to fostering students' deep understanding of the conceptual components of a specific learning domain and less to communicating the goals of scientific, often domain-general, methods and tools (Chase and Klahr 2017). Thus, in order to promote students' knowledge related to scientific methods, DI approaches either may be as effective as PF approaches (cf., Chase and Klahr 2017) or even be more effective (as indicated by our first study). Hence, the effectiveness of PF may depend not only on the structuredness of the learning topic (cf., Loibl et al. 2017), but also on the type of knowledge that is addressed. The OECD (2006), for instance, distinguishes between students' knowledge of science (i.e., understanding of scientific concepts and theories) and their knowledge about science (i.e., understanding of the means and goals of science). Given the argument developed by Chase and Klahr (2017), one could hypothesize that the benefits of PF rather apply to students' acquisition of knowledge of science than to their acquisition of knowledge about science (i.e., about natural *and* social sciences).

To clarify, all of the aforementioned explanatory conjectures require more research to examine them further. However, the findings of our two studies demonstrated that the PF effect did not transfer from learning in STEM domains to learning social science research methods, although crucial features of the PF design were fulfilled. Our results further suggest that this non-effect of PF may relate to PF students' lack of failure awareness prior to instruction and a missing increase of their perceived competence during instruction. So far, PF students' awareness of their failure or their competence limitations *prior to instruction* have not been assumed to be relevant for the PF effect to occur. PF students' development of perceived competence during instruction has not yet even been considered being a potential mechanism of the PF effect. Above, we provide theoretical and partly also empirically supported conjectures on why these mechanisms could be important and how they could be promoted. Nevertheless, (further) empirical

investigation is needed to determine whether our explanatory conjectures are applicable, whether students' awareness of their competence limitations prior to instruction and their development of perceived competence during instruction are indeed mechanisms underlying the PF effect, and whether the benefits of PF are affected by, for instance, designing the PF problem in such a way that students receive feedback through their solutions.

Limitations

The present studies were conducted in an out-of-school lab for social sciences. As such labs are usually visited by whole classes, we had to randomly assign the students on a class level to the PF and DI condition. Although quasi-experimental designs promote the external validity of investigations (especially in the context of educational research), they also bear the risk of non-equivalent test groups due to the lack of randomization of individual participants. To address this limitation, we conducted multilevel regression analyses, which account for the clustered structure of our data (i.e., individual students clustered within whole classes). Due to the small number of clusters ($N=11$ in study 1 and $N=7$ in study 2), the multilevel models only serve as analysis in comparison and addition to our main analyses (cf., Cress 2008). With respect to the effect of condition on students' learning outcome, the results of the multilevel analyses (calculated in Mplus 7.31) are in line with our findings from the ANCOVAs, i.e., a negative effect of PF on students' learning of social science research methods in study 1 ($\beta=-3.07$, $SE=0.64$, $p<0.001$, $ICC=0.11$) and no effect of PF in study 2 ($\beta=-0.25$, $SE=0.69$, $p=0.71$, $ICC=0.01$).

A further limitation relates to the interpretation of the aforementioned non-effect of our experimental manipulation in study 2. More specifically, a null hypothesis significance test, such as the *F*-test, examines whether a null hypothesis can be rejected (which was in line with our expectation) and not whether a null hypothesis can be accepted (which is contrary to our hypothesis but indicated by the results of study 2). Hence, non-effects revealed by null hypothesis significance tests are difficult to interpret. In line with the suggestions by Aberson (2002), we addressed this difficulty in the results sections above by calculating and reporting the confidence intervals of the effect sizes. In addition, we conducted a Bayesian variance analysis in order to examine the effect of condition on students' learning in study 2. The Bayesian ANOVA (calculated in JASP 0.9.2.0) revealed moderate evidence for the null hypothesis ($BF_{10}=0.23$; see Wagenmakers et al. (2018) for the interpretation of the Bayes factor), thus supporting the results of our previous ANCOVA.

Another limitation is that we were unable to use a pretest to assess students' knowledge about social science research methods. To avoid an activation of PF students' prior knowledge about principles of experimental design in study 1 and the differences between correlational and causal evidence in study 2 prior to the problem-solving phase, we did not assess students' actual understanding of the targeted learning domain prior to the studies. This difficulty is also common in previous PF studies. For instance, Loibl and Rummel (e.g. 2014a) used a pretest that only assessed students' prior knowledge relevant for generating the intuitive solution ideas during the initial problem-solving phase, but not their knowledge on the targeted learning concept. Students' performance on this pretest, however, did not correlate with their learning outcome (Loibl and Rummel 2014a). Newman and DeCaro (2018) even assumed that the pretest in their first study caused the non-effect of PF. In their second study, the authors did not assess students' prior knowledge in

a pretest, and the PF effect was demonstrated. Hence, we decided not to assess students' prior knowledge in a pretest.

One could further argue that the design of our tasks constitutes a further limitation, as our tasks did not allow PF students to receive feedback through their generated solutions and in turn to become aware of their failure during the problem-solving phase, as was seemingly the case in previous classical PF studies on learning in mathematics. However, rather than being a limitation, we believe that our studies promote a research agenda investigating the role played by task design in the effectiveness of PF. Specifically, so far, the PF literature has not explicitly stated that the task has to be designed in such a way that students' awareness of their failure is fostered during problem solving. Instead, it has only been emphasized that the PF problem has to meet an appropriate level of complexity, such that students are hindered from solving the problem canonically and thus fail during the initial problem-solving phase (Kapur and Bielaczyc, 2012). As our results with respect to design-related question 1 demonstrated, the tasks in the present two studies fulfilled this design principle of PF.

Conclusion

To conclude, in line with most previous studies testing the effect of problem solving prior to instruction on learning in a domain other than mathematics or science (see Table 2), our two studies did not replicate the effectiveness of PF for learning social science research methods. These findings are particularly interesting in light of the fact that, in contrast to previous studies on the PF effect on learning in non-STEM domains, our studies emulated the typical design features of classical PF studies. Although our findings do not offer conclusive evidence why the PF approach was not more effective than DI for learning social science research methods, the results do have interesting implications for future research. We have learned that the two major features of the PF design—asking students to invent solutions to a novel and complex problem during an initial problem-solving phase and explaining the canonical solution to students by comparing and contrasting typical erroneous solutions with the canonical solution during a final instruction phase—are not sufficient to evoke the PF effect. Students' activation of prior knowledge during problem solving and their awareness of knowledge limitations during instruction are also not enough to enable successful learning through PF. However, our results suggest that it might be important for the effectiveness of PF that students struggle and become aware of their competence limitations during the initial problem-solving phase and, in turn, that they develop perceived competence during instruction. Altogether, the failure to replicate the effect of PF on learning social science research methods suggests new directions for future research examining what makes PF effective: To date, PF literature has not explicitly stated that students need to become aware of their failure or competence limitations during the initial problem-solving phase, or that the PF design has to foster this awareness. The seminal paper on PF design principles by Kapur and Bielaczyc (2012) does not describe PF students' awareness of their failure during initial problem solving as one of the design principles. Kapur and Bielaczyc (2012) only emphasize the importance of an appropriate level of complexity of the problem-solving task, which should be neither too high (such that students cannot generate any solution ideas) nor too low (such that students can solve the problem canonically). Thus, they point to the relevance of the process of failing (which

was fulfilled in our two studies), but not to the process of becoming aware of one's failure during the initial problem-solving phase. In contrast, Loibl et al. (2017) explicitly describe students' awareness of knowledge gaps as one of three mechanisms that may be important for the PF effect to emerge. However, they also emphasize that complex and rich problems as implemented in PF do not easily allow students to evaluate their solution ideas and thus to become aware of their knowledge gaps during the initial problem-solving phase. Therefore, Loibl et al. (2017) state that using erroneous student solutions during instruction (also fulfilled in our two studies) is crucial for fostering students' awareness of knowledge gaps and thereby promoting the effectiveness of PF for learning. Hence, as our findings suggest that students' awareness of failure prior to instruction may be crucial for the effectiveness of PF, our studies provide an important and novel contribution to research on PF.

Acknowledgements The studies were funded by the Graduate School of Educational Studies of Ruhr-University Bochum (RUB) within the doctoral program on science education in out-of-school labs. We thank the members of that program for their suggestions on the design of these studies. We would also like to thank University of Washington Professor Philip Bell and his research group for their reviews on drafts of this article. We thank the German-American Fulbright Commission for funding a research stay in Philip Bell's research group at the University of Washington in Seattle. We are very thankful to the participating schools and teachers for their organizational efforts, and to the team of the Alfried Krupp-Schülerlabor (out-of-school lab) for their cooperation. We thank our student research assistants and interns for their help in collecting and coding the data. Last, but definitely not least, we want to acknowledge the invaluable support that we received while writing this paper from our colleagues at the Educational Psychology Research Group at RUB.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Data collection followed the standards of the German Psychological Society (DGPs). Given the type of data collected (i.e., anonymous data), no ethical approval was considered necessary according to the guidelines at that time.

Informed consent Written informed parental consent was given for all participants.

References

Aberson, C. (2002). Interpreting null results: Improving presentation and conclusions with confidence intervals. *Journal of Articles in Support of the Null Hypothesis, 1*(3), 36–42.

Chalmers, A. F. (2013). *What is this thing called science?* Indianapolis, IN: Hackett Publishing.

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: For some content, it's a tie. *Journal of Science Education and Technology, 26*(6), 582–596.
Cho, Y. H., Caleon, I. S., & Kapur, M. (2015). *Authentic problem solving and learning in the 21st century*. Singapore: Springer.
Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.
Cress, U. (2008). The need for considering multilevel analysis in CSCL research—An appeal for the use of more advanced statistical methods. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 69–84.
Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development, 66,* 1101–1118.
De Bruin, A. B. H., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education, 51*(6), 575–584.
De Bruin, A. B., & van Merriënboer, J. J. G. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51,* 1–9.
Euler, M. (2004). Quality development: Challenges to physics education. In M. Michelini (Ed.), *Quality development in teacher education and training* (pp. 17–30). Udine: Forum.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.
Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. *British journal of educational psychology, 84*(3), 502–519.
Garner, N., & Eilks, I. (2015). The expectations of teachers and students who visit a non-formal student chemistry laboratory. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(5), 1197–1210.
Gee, J. P. (2005a). Good video games and good learning. *Phi Kappa Phi Forum, 85,* 33–37.
Gee, J. P. (2005b). Learning by design: Good video games as learning machines. *E-Learning and Digital Media, 2*(1), 5–16.
Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction, 39,* 72–87.
Glowinski, I., & Bayrhuber, H. (2011). Student labs on a university campus as a type of out-of-school learning environment: Assessing the potential to promote students' interest in science. *International Journal of Environmental and Science Education, 6*(4), 371–392.
Hmelo-Silver, C. E., & Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-Based Learning, 1*(1), 21–39.
Jacobson, M. J., Kim, B., Pathak, S., & Zhang, B. (2015). To guide or not to guide: Issues in the sequencing of pedagogical structure in computational model-based learning. *Interactive Learning Environments, 23*(6), 715–730.
Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development, 45*(1), 65–94.
Kagan, J. (2009). *The three cultures: Natural sciences, social sciences, and the humanities in the 21st century*. New York: Cambridge University Press.
Kant, J. M., Scheiter, K., & Oschatz, K. (2017). How to sequence video modeling examples and inquiry tasks to foster scientific reasoning. *Learning and Instruction, 52,* 46–58.
Kapur, M. (2008). Productive failure. *Cognition and instruction, 26*(3), 379–424.
Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523–550.
Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579.
Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*(4), 651–672.
Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022.
Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice, 1*(1), 51–65.
Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299.

Kapur, M., & Bielaczyc, K. (2011). Classroom-based experiments in productive failure. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2812–2817). Austin, TX: Cognitive Science Society.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *The Journal of the Learning Sciences, 21*(1), 45–83.

Kapur, M., & Lee, J. (2009). Designing for productive failure in mathematical problem solving. In N. Taatgen, & V. R. Hedderick (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2632–2637). Austin, TX: Cognitive Science Society.

Kapur, M., & Rummel, N. (2009). The assistance dilemma in CSCL. In A. Dimitracopoulou, C. O'Malley, D. Suthers, & P. Reimann (Eds.), *Computer supported collaborative learning practices—CSCL 2009 Conference Proceedings* (Vol. 2, pp. 37–42). Berlin: International Society of the Learning Sciences.

Kapur, M., & Toh, L. (2015). Learning from productive failure. In Y. H. Cho, I. S. Caleon, & M. Kapur (Eds.), *Authentic problem solving and learning in the 21st century* (pp. 213–227). Singapore: Springer.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological science, 15*(10), 661–667.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it. Delaying instruction improves mathematics problem solving: A classroom study. *The Journal of Problem Solving, 7*(1), 36–49.

Loibl, K., & Rummel, N. (2014a). Knowing what you don't know makes failure productive. *Learning and Instruction, 34,* 74–85.

Loibl, K., & Rummel, N. (2014b). The impact of guidance during problem solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*(3), 305–326.

Loibl, K., & Rummel, N. (2015). Productive failure as strategy against the double curse of incompetence. *Learning: Research and Practice, 1*(2), 113–121.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review, 29*(4), 693–715.

Marei, H. F., Donkers, J., Al-Eraky, M. M., & van Merrienboer, J. J. (2017). The effectiveness of sequencing virtual patients with lectures in a deductive or inductive learning approach. *Medical teacher, 39*(12), 1268–1274.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science, 41*(3), 621–634.

Mazziotti, C., Rummel, N., Deiglmayr, A., & Loibl, K. (2019). Probing boundary conditions of Productive Failure and analyzing the role of young students' collaboration. *Science of Learning, 4*(2), 1–9.

Nachtigall, V., Rummel, N., & Serova, K. (2018). Authentisch ist nicht gleich authentisch–Wie Schülerinnen und Schüler die Authentizität von Lernaktivitäten im Schülerlabor einschätzen. [Authentic does not equal authentic - how students evaluate the authenticity of learning activities in an out-of-school lab]. *Unterrichtswissenschaft, 46*(3), 299–319.

Newman, P., & DeCaro, M. (2018). How much support is optimal during exploratory learning? In *Proceedings of the 40th annual conference of the cognitive science society.* Madison, WI: Cognitive Science Society.

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology, 22*(3), 453–460.

OECD (Organisation for Economic Co-operation and Development). (2001). *Knowledge and skills for life. First results from the OECD programme for international students Assessment (PISA) 2000.* Paris: OECD.

OECD (Organisation for Economic Co-operation and Development). (2006). *Assessing scientific, reading and mathematical literacy A framework for PISA 2006*. Paris: OECD.

Pauly, Y. (2012). Was sind und zu welchem Zweck brauchen wir geisteswissenschaftliche Schülerlabore? [What are and for what purpose do we need out-of-school labs on humanities and social sciences]. *Handbuch Wissenschaftskommunikation* (pp. 205–210). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175–189.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of educational psychology, 93*(2), 346–362.

Scharfenberg, F. J., & Bogner, F. X. (2014). Outreach science education: Evidence-based studies in a gene technology lab. *Eurasia Journal of Mathematics, Science & Technology Education, 10*(4), 329–341.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and instruction, 16*(4), 475–522.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184.

Sinha, T., & Kapur, M. (2019). When productive failure fails. In *Proceedings of the 41th annual conference of the cognitive science society.* Montreal: Cognitive Science Society.

Stevens, R. J., Slavin, R. E., & Farnish, A. M. (1991). The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology, 83*(1), 8–16.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

Tam, K. (2017). *Examining productive failure instruction in dental ethics.* (Unpublished doctoral dissertation). The University of Arizona, Tucson, AZ,

Tawfik, A. A., Rong, H., & Choi, I. (2015). Failing to learn: towards a unified design approach for failure-based learning. *Educational Technology Research and Development, 63*(6), 975–994.

VanLehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 19–41). New York: Springer.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., …, Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76.

Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock.' *European Educational Research Journal, 8*(3), 476–483.

Weaver, J. P., Chastain, R. J., DeCaro, D. A., & DeCaro, M. S. (2018). Reverse the routine: Problem solving before instruction improves conceptual knowledge in undergraduate physics. *Contemporary Educational Psychology, 52,* 36–47.

Westermann, K., & Rummel, N. (2012). Delaying instruction: Evidence from a study in a university relearning setting. *Instructional Science, 40*(4), 673–689.

Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzskala intrinsischer Motivation (KIM) [Testing a short scale of intrinsic motivation]. *Zeitschrift für Didaktik der Naturwissenschaften, 15,* 31–45.