Teaching and Teacher Education 109 (2022) 103518

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Research paper

Professional development for language support in science classrooms: Evaluating effects for elementary school teachers

Birgit Heppt [a, *], Sofie Henschel [b], Ilonca Hardy [c], Rosa Hettmannsperger-Lippolt [c, 1], Katrin Gabler [a, 2], Christine Sontag [a, 2], Susanne Mannel [c, 3], Petra Stanat [b]

[a] Humboldt-Universität zu Berlin, Germany
[b] Institute for Educational Quality Improvement (IQB) at Humboldt-Universität zu Berlin, Germany
[c] Goethe University Frankfurt, Germany

HIGHLIGHTS

- We evaluated a professional development program (PD) in a quasi-experimental field trial.
- The PD included courses on elementary school science topics and on subject-integrated language support.
- We found strong treatment effects on elementary school teachers' language-support skills.
- Effects on language-support activities in the classroom were smaller and differed across science topics.
- For science, positive effects emerged for one out of two topics and for teachers' self-efficacy.

ARTICLE INFO

Article history:
Received 1 February 2021
Received in revised form
1 September 2021
Accepted 11 September 2021
Available online 21 October 2021

Keywords:
Professional development
Language support
Elementary school
Science education

ABSTRACT

We investigate the effectiveness of professional development (PD) aimed at promoting teachers' language-support skills in elementary school science instruction. In a 2-year quasi-experimental field trial study with 32 teachers in Germany, an intervention group (IG) and a control group (CG) received PD for teaching selected science topics; the IG additionally received PD for language support. Strong treatment effects emerged on teachers' language-support skills and, to a lesser extent, on language support activities in classroom teaching. All teachers gained pedagogical content knowledge and self-efficacy for teaching elementary school science, thus pointing to the effectiveness of the PD.

1. Introduction

To help students develop the language skills needed for mastering the language demands of schooling, teachers are increasingly expected to integrate language support into their lessons in all subjects. Yet, findings of various OECD countries, including Germany, indicate that teachers often do not feel well-equipped to implement language support in their regular classes (Bunch, 2013; Lucas et al., 2008; Quílez, 2021; Romijn et al., 2021; Thomassen & Munthe, 2020). Since mandatory courses on language support and second language acquisition are still an exception in teacher education programs (Paetsch & Heppt, in press), teachers need to have access to effective professional development (PD) for acquiring the skills required for integrating language-support into subject-specific teaching.

Considerable efforts have been undertaken in various countries to develop and implement PD aimed at improving teachers' language-support skills (for an overview, see Bunch, 2013; Schneider et al., 2012). While most of these programs prove effective to some extent, there is a huge variability in PD designs,

* Corresponding author. Correspondence concerning this article should be addressed to Birgit Heppt, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099, Berlin, Germany.

E-mail address: birgit.heppt@hu-berlin.de (B. Heppt).

[1] Rosa Hettmannsperger-Lippolt is now at Hessische Lehrkräfteakademie, Walter-Hallstein-Straße 3–7, 65197 Wiesbaden, Germany.

[2] Katrin Gabler and Christine Sontag are now at Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany.

[3] Susanne Mannel is now at accadis Hochschule Bad Homburg, University of Applied Sciences, Am Weidenring 4, 61352 Bad Homburg, Germany.

https://doi.org/10.1016/j.tate.2021.103518
0742-051X/

targeted language-support approaches, and research methodology with only a few studies implementing experimental trials (Kalinowski et al., 2019, 2020). Moreover, only a small number of studies deliberately aimed at promoting both language-support skills *and* content knowledge (Hart & Lee, 2003; Lee & Maerten-Rivera, 2012). Against this background, the present study investigates the effectiveness of PD aimed at improving teachers' knowledge and skills for integrating language-support strategies into their regular elementary school science instruction. To facilitate the integration of language support into subject-specific science teaching and to ensure comparability across classrooms, the program also included PD on teaching selected elementary school science topics.

1.1. Promoting students' language skills in science classes

Inquiry-based and constructivist science classes that engage students in active and collaborative learning, provide ample opportunities for using and developing language in meaningful contexts while co-constructing knowledge (Furtak et al., 2012; Stoddart et al., 2002). Such science classes are therefore optimal learning environments for integrating content and language-learning (cf. Dawes, 2004; Fang & Wei, 2010; Stoddart et al., 2002).

Science classes typically include activities of scientific inquiry to foster students' scientific argumentation skills and understanding of how scientific knowledge is generated and refined (e.g., Hardy et al., 2010; Lee & Maerten-Rivera, 2012). Typical steps included in scientific inquiry are the formulation of hypotheses, the planning and implementation of experiments, and the interpretation of results (Abd-El-Khalick et al., 2004; Dawes, 2004; Hardy et al., 2010; Minner et al., 2010). These activities clearly aim at advancing conceptual understanding and content knowledge, but they also require the mastery of the language register of schooling (i.e., academic language; cf. Ødegaard et al., 2014; Seah & Silver, 2018). To explain the results of an experiment, for instance, students need to be able to report from a rather distanced point of view, they need to know the correct technical terms (e.g., “buoyancy force”, “water displacement”), and they need to be aware of the connectives and grammatical structures that are specific to the language function of explaining (e.g., “because”, “due to”; Oyoo, 2011; Quílez, 2021). Besides these communicative purposes, language is also an important tool for knowledge construction (Gentner, 2016; Hardy et al., 2020; Studhalter et al., 2021). Vocabulary knowledge, for example, serves as a basis for acquiring new concepts or restructuring preconcepts, as it enables cognitive operations such as comparisons and logical inferences (Gelman & Markman, 1986; Gentner, 2016).

Delivering didactically sophisticated science classes, which foster both students' conceptual understanding and their language proficiency, is highly challenging. Teachers need substantial content knowledge (CK) and pedagogical content knowledge (PCK; Shulman, 1986) in the respective subject(s) to provide effective inquiry-based science classes. Teachers' PCK has been identified as particularly important for students' (self-assessed) science competence (Lange et al., 2012; Sadler et al., 2013) and interest (Fauth et al., 2019; Lange et al., 2012).

Moreover, teachers need to be equipped with skills and strategies of language support for initiating meaningful and effective classroom discourse and learning activities. One such approach is linguistic scaffolding (Gibbons, 2002). Drawing on basic principles of Vygotsky's theory of social learning (Wood et al., 1976), this approach aims at moving students from colloquial everyday language to a more elaborate use of academic language (cf. Darsow et al., 2012; Gibbons, 2002; Lucero, 2014; Mahan, 2020). In a typical sequence, students employ their everyday language skills in the language of instruction during learning activities, allowing them to acquire a basic understanding of the targeted concepts. With specific linguistic aids (scaffolds) implemented in the following learning activities, they successively develop their language skills towards a more sophisticated and precise language use (i.e., academic language) and deepen their conceptual content knowledge. In order to facilitate students' language learning, teachers use a variety of language-support strategies. These pertain to teachers' language input (e.g., acting as language role models by frequent use of academic language), use of questions (e.g., asking open-ended questions to initiate more elaborate student utterances), language feedback (e.g., rephrasing students' utterances using the academic language register), and strategies for actively shifting students' attention (e.g., by using visual aids; for further examples see Table 1; Darsow et al., 2012; Gabler et al., 2020; Mahan, 2020; Wasik, 2010). The linguistic aids are gradually removed, adapting to the students' current state of language proficiency.

1.2. Characteristics of effective teacher PD

A growing body of research on the effectiveness of teacher PD overall (e.g., Darling-Hammond et al., 2009; Garet et al., 2001; Guskey & Yoon, 2009; Lipowsky & Rzejak, 2015; Sancar et al., 2021; Timperley et al., 2007) and studies with a specific focus on PD for language support across the curriculum (for an overview, see Kalinowski et al., 2020; Kalinowski et al., 2019) yielded convergent findings regarding the key characteristics of effective teacher PD. Much of the literature suggests that effective PD has a narrow focus on subject-specific knowledge and uses a variety of formats (e.g., workshops, coaching, online discussions) that are delivered in multiple sessions over an extended period of time (e.g., Darling-Hammond et al., 2009; Garet et al., 2001; Sancar et al., 2021). However, there is no simple link between the duration of a PD and its effectiveness (Guskey & Yoon, 2009; Timperley et al., 2007). Longer and more varied PDs seem to be more effective, as they are more likely to combine phases of input with practice and reflection (Darling-Hammond et al., 2009; Lipowsky & Rzejak, 2015).

During application phases, teachers typically implement the newly acquired knowledge into their classroom teaching. These active implementation phases should include feedback and foster reflection (Ingvarson et al., 2005; Lipowsky & Rzejak, 2015; Romijn et al., 2021). An effective way to stimulate reflection on and development of teaching practices is the analysis of video-recorded lessons (e.g., Blomberg et al., 2014; Fukkink & Kramer, 2011; Piwowar et al., 2013).

Collaboration among participants forms another important aspect of effective PD (e.g., Armour & Yelling, 2007; Babinski et al., 2018; Garet et al., 2001; Geldenhuys & Oosthuizen, 2015; Sancar et al., 2021). It can be promoted by engaging teachers in discussions, group work, and joint reflections of newly acquired knowledge and practical experiences, for instance (for an overview, see Kalinowski et al., 2019).

1.3. Effectiveness of teacher PD for language support across the curriculum

Typically, evaluation studies assess the extent to which PD is effective, focusing on one or several conceptual levels: teachers' immediate reaction to the program, i.e., their satisfaction with and acceptance of the PD intervention (Level 1); teachers' learning gains and intended changes in knowledge and motivational orientations (Level 2); changes in teachers' classroom behaviors and practices (Level 3); and students' learning gains and competence development (Level 4), presenting the ultimate goal of teacher PD

Table 1
Typical language-support strategies within the scaffolding approach (e.g., Gabler et al., 2020).

Input	Questions	Feedback	Attention Focus
Acting as language role model • Giving well-considered and elaborate language input (target word "observe/observation": "I am interested in your observations. Please describe in detail what you observed during the experiment.") • Mapping one's own or students' actions with language (e.g., "First I pour water into the big cup. Then I press the small cup", while demonstrating exactly this action to the students.) • Using thinking aloud-techniques	Language-supportive questions • Asking open-ended questions ("What would you expect and why?") • Asking questions with a specific focus on language ("Which other word do we know for this?")	Correcting and extending students' utterances • Giving corrections in terms of content, vocabulary, and grammar • Rephrasing (and extending), using the academic language register ("The liquid made itself bigger" – "Exactly. The liquid expanded in the thermometer.")	Explicitly drawing students' attention to terms and phrases • Giving oral hints ("What would a scientist call this?") • Prompting students to self-corrections • Using materials that facilitate comprehension (e.g., word cards, printing new words in bold in a text) • Conducting explicit exercises (e.g., when weighing objects of different sizes and material kind, students can be prompted to formulate comparisons using expressions like "as heavy as", "heavier" etc.)

(Guskey, 2000; Lipowsky & Rzejak, 2015). Given the scope of the present investigation, our literature review focuses on studies evaluating the effects of teacher PD for language support at Levels 2 and 3, namely teachers' cognition and classroom practices (for studies, evaluating PD effects at the student level, see August et al., 2014; Babinski et al., 2018; Brisk & Zisselsberger, 2011; Kalinowski et al., 2019).

As for cognition, teacher PD may impact various dimensions, such as teachers' self-efficacy to use language support in instruction as well as their beliefs and knowledge, all of which form part of teachers' professional competence (Kunter et al., 2013). In their recent meta-analysis, Kalinowski et al. (2020) found a small and non-significant training effect on teachers' cognition across studies ($g' = 0.21$, $SE = 0.14$). Yet, only four of the ten studies that were included in the meta-analysis reported cognitive measures (e.g., knowledge, beliefs, self-efficacy). Whereas two studies focused on teachers' self-reported knowledge, there were no studies that evaluated actual knowledge gains, using objective tests. Lee and Maerten-Rivera (2012), for instance, conducted a 3-year teacher PD in the US, aimed at improving teachers' science instruction and support of English language learners' (ELLs) literacy development. The authors found gains in teachers' self-reported science knowledge and, to a lesser extent, changes in the self-reported implementation of teaching practices aimed at ELLs' language development.

While teachers' (objectively measured) knowledge and skills were rarely assessed in previous research, the majority of studies investigated teachers' classroom practice upon PD completion. Kalinowski et al. (2020) report an overall medium to large effect ($g' = 0.71$, $SE = 0.16$) and conclude that teacher PD seems to improve teachers' classroom practice. Yet, the studies included in the meta-analysis varied considerably in terms of language-support approaches implemented in the teacher PDs, the assessments of teachers' classroom practices, and the resulting effects. In a study by Hart and Lee (2003), teachers participated in several workshops aimed at improving their engagement of students in science inquiry while integrating English language and literacy in their science classes. After the 1-year intervention, teachers provided more linguistic scaffolding, resulting in a small effect size ($d = 0.29$). Instructional strategies aimed at promoting literacy activities (i.e., reading and writing) during science classes, however, did not change ($d = 0.08$).

One of the very few studies conducted outside the US focused on kindergarten teachers' use of scientific reasoning (e.g., "comparing", "explaining") and domain-specific academic vocabulary (e.g., "air", "press") in early science instruction in the Netherlands (Henrichs & Leseman, 2014). The authors found that, after a 3-h training on academic language, teachers in the IG asked more questions aimed at enhancing students' scientific reasoning and used more domain-specific academic vocabulary than teachers in the CG. These findings are mirrored by a recent study, also conducted in the Netherlands, that aimed at increasing elementary school teachers' use of open-ended questions and language-learning strategies during science classes (van Dijk et al., 2019). This study implemented one-on-one video-based coaching. After the intervention, teachers from the IG used more open-ended questions in their science classes and their oral communication increased in complexity and sophistication.

Most of the above studies on PD effects on teachers' classroom practices used instruments that were developed for the purpose of the particular investigation. Yet, two recent studies from the US included standardized measures that had been used and validated before (Babinski et al., 2018; Tong et al., 2018). Based on an established classroom observation tool and a researcher-developed protocol, Babinski et al. (2018) evaluated the effects of PD with a specific focus on language support for ELLs. Whereas the IG and the CG showed large posttest differences on the researcher-developed measure, few differences emerged on the established tool. These results are in line with findings by Kalinowski et al. (2020) who identified the degree of standardization of an instrument as a potential moderator of the PD's effectiveness. Evaluation studies that rely exclusively on researcher-developed instruments thus tend to report larger treatment effects than studies that include well-established, standardized measures.

In sum, the emerging literature on PD effects indicates that teacher PD for subject-integrated language support may be beneficial for some aspects related to teachers' cognition, specifically beliefs and self-reported knowledge (e.g., Hart & Lee, 2003; Kalinowski et al., 2020). Effects on teachers' classroom practice have been studied more extensively than cognitive aspects, and results point to larger effects. Despite these promising findings, research in this field is still at an early stage. Previous research has largely neglected teachers' learning gains as reflected by objective performance measures and there is a strong focus on researcher-developed measures instead of validated and well-established measures. The vast majority of studies was conducted in the US and almost all of these investigations aimed at training teachers to foster language skills of ELLs. These are students with limited English proficiency as assessed by standardized tests, among them many students with an immigrant background. Immigrant students are also an important target group for language support in Germany (e.g., Paetsch et al., 2014; Stanat et al., 2012), but there is a growing body of research showing that academic language is challenging for both multilingual and monolingual students (Heppt et al., 2016; Prediger & Zindel, 2017). Current educational policies and curriculum developments in Germany therefore call for subject-integrated academic language support for *all* students (cf. Becker-Mrotzek & Roth, 2017). Thus, teacher trainings should improve teachers' skills to integrate language-support strategies into their regular domain-specific teaching to support all students' development of content knowledge and language skills.

2. The present study

The goal of the present study is to examine the extent to which a PD program based on the scaffolding approach (Gibbons, 2002) fosters in-service teachers' skills to provide (academic) language support in elementary school science instruction. We trained participants from both CG and IG for teaching selected elementary school science topics. The IG additionally received extensive training for incorporating language-support strategies into their science teaching. Our PD evaluation focuses on treatment effects at Levels 2 and 3, that is, effects on teachers' cognition and classroom practices, and includes both language-support skills *and* elementary school science teaching.

In terms of cognition measures, we focus on knowledge gains, pertaining to both participants' language-support skills and their PCK on selected elementary school science topics. We additionally investigate changes in teachers' self-efficacy for teaching elementary school science. Teachers' self-efficacy is an important driver of subsequent behavior and teaching effectiveness (Klassen & Tze, 2014) and can be increased by subject-matter training (Cantrell & Hughes, 2008; Mulholland et al., 2004).

Regarding teachers' classroom practices, we investigate three dimensions of instructional support, as measured by the Classroom Assessment Scoring System (CLASS K-3; Pianta et al., 2008). "Language modeling" has a clear focus on teachers' language-support practices and entails a range of strategies IG-teachers learned and applied in our PD (Table 1). "Concept development" and "quality of feedback", however, are more general dimensions of instructional quality, aimed at improving students' content knowledge and conceptual understanding.

2.1. Research questions and hypotheses

Specifically, we explored the following research questions and hypotheses:

(1) Do IG-teachers differ from CG-teachers in their language-support skills (cognition; Level 2) and in their classroom teaching (practice; Level 3) after completing the PD for language support in science classrooms? We expected the IG to increase their language-support skills and to outperform the CG upon PD completion. With regard to teachers' instructional support in classroom teaching, we assumed advantages of the IG over the CG for language modeling but not for concept development. As the CLASS-dimension quality of feedback broadly covers feedback strategies with no specific focus on language support, it may be assumed that IG and CG perform equally well. Yet, feedback strategies are also a common means for fostering students' language skills and were promoted in the PD for language-support (e.g., by correcting and expanding students' utterances or by rephrasing responses, using more specific and elaborate vocabulary; see Table 1; e.g., Dannenbauer, 2002; Gabler et al., 2020; Lyster & Saito, 2010). Given these somewhat contradictory assumptions, we did not formulate a hypothesis for quality of feedback.

(2) Do IG and CG-teachers increase their PCK and their self-efficacy (cognition; Level 2) for teaching elementary school science after completing the PDs for teaching selected elementary school science topics? The PDs had a strong focus on subject-specific knowledge and encouraged participants' social interaction and collaborative learning, the latter of which is regarded as particularly important for boosting self-efficacy (Bandura, 1997; Yang, 2020). We therefore hypothesized that all participants would gain in subject-specific PCK and increase their self-efficacy for teaching elementary school science.

3. Method

3.1. Study design and data collection

We conducted a field trial study to evaluate the effects of our PD program (Fig. 1). Experimental designs with at least two groups are generally seen as a highly appropriate way for testing treatment effects (Sullivan, 2011). Our quasi-experimental design therefore involved an IG and a (waiting) CG. The study took place over the course of two school years and consisted of two major phases: During the PD phase in Year 1, both IG and CG participated in three face-to-face PD courses on teaching the topics of "floating and sinking" (Topic 1), "evaporation and condensation" (Topic 2), and "education for sustainable development" (Topic 3). The IG additionally received extensive training for fostering students' language skills during regular science instruction and participated in coaching sessions in conjunction with a number of classroom trials. The face-to-face training on language support was based on the science topic of "floating and sinking" and was therefore conducted after the course on Topic 1. During the implementation phase in Year 2, teachers taught the three science topics in their regular science classes in Grades 3 and 4. Given the time lag between PD and implementation, the present study measured delayed effects on teachers' practice.

To evaluate PD effects, we administered written assignments prior to the beginning of the PD program, before and after completion of each science unit, and after the training series on language-support skills (Fig. 1). During the implementation phase, we videotaped the second double lesson (90 min) of each science topic in all participating classrooms. We had planned to videotape three double lessons per teacher, one in each of the three different science topics. However, due to sample attrition throughout the project, only data on the first two topics could be used for evaluating PD effects in the present article. After completing the implementation phase, the CG had the chance to participate in a shortened and optimized version of the PD for language support in science classrooms.

3.2. Sample selection and sample

The current study was conducted in two federal states of Germany. To recruit teachers for participation in the study, we contacted elementary schools and carried out introductory events for school principals and teachers. We aimed at avoiding spill-over effects from IG to CG (e.g., with teachers from different treatment groups exchanging PD materials or engaging in joint lesson planning), as this might have familiarized CG-teachers with the content of PD for language support and, thus, blurred treatment effects. Teachers within one school were therefore assigned to the same treatment condition.

Participation in the study was voluntary for schools, teachers, and students. The project was very time-consuming for teachers, especially for those in the IG, and required their ongoing participation for more than 2 years. We therefore offered participants several incentives at different stages of the study (e.g., teaching materials, a certificate for their participation in the PD program,

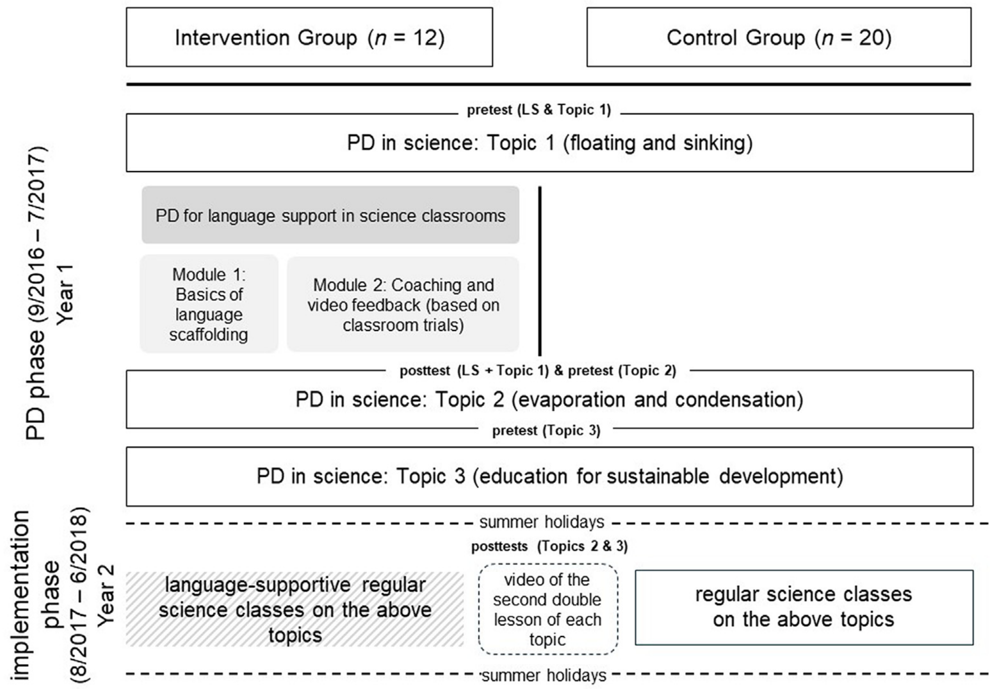

Fig. 1. Overview of study design and data collection. PD = professional development. LS = language support.
Note. The subject-related posttest for Topics 2 and 3 took place shortly before the respective teaching units. Due to sample attrition throughout the project, only Topics 1 and 2 were included in the evaluation. PD for language support for the (waiting) control group took place after the implementation phase (9/2018-12/2018).

financial allowances for the class). Nevertheless, the project suffered from substantial sample attrition (see section "Sample Attrition"). Out of 44 teachers who participated in the pretest, 32 teachers (26 female) from 16 elementary schools completed the PD program and took part in at least one posttest. In the following, these teachers are referred to as "cognition sample" and serve as the basis for evaluating PD effects at Level 2. Of these 32 teachers, 24 further participated in the project during the implementation phase. This group, which is a strict subsample of the cognition sample, is referred to as "practice sample" and is used for evaluating the effectiveness of the PD at Level 3.

Table 2 displays basic demographics for the cognition sample ($n_{IG} = 12$, $n_{CG} = 20$) at T1; the respective data for the practice sample ($n_{IG} = 7$, $n_{CG} = 17$) are shown in Table 3. The following details refer to the cognition sample, but a very similar pattern of results emerged for the practice sample. The two treatment groups in the cognition sample did not differ in terms of age ($t = 0.16$, $df = 24$, $p = .87$, $d = 0.07$, 95% CI [-0.80, 0.94]) and gender ($\chi^2(1) = 0.6$, $p = .82$, $\varphi < .04$, 95% CI [.00, .36]) and the gender distribution largely corresponds to the numbers reported for

Table 2
Demographic data for the intervention group and the control group at T1: Cognition sample.

Variable	Intervention Group (n = 12 from 7 schools)			Control Group (n = 20 from 9 schools)		
	n (%)[a]	*M*	*SD*	*n* (%)[a]	*M*	*SD*
Gender						
Female	10 (83.30)			16 (80.00)		
Male	2 (16.70)			4 (20.00)		
Federal state						
State 1	12 (100)			2 (10.00)		
State 2	-- (0)			18 (90.00)		
Mean age in years		41.14	9.33		41.74	7.96
Background in elementary school education	8 (72.70)			19 (95.00)		
Background in elementary school science	2 (18.20)			9 (45.00)		
Background in language support/German as a second language	6 (54.50)			3 (16.70)		

Note. The cognition sample includes all teachers who participated in the project throughout the entire professional development phase and took part in the written assignments of the pretests and posttests.
[a] Information is based on the valid percentages.

Table 3
Demographic data for the intervention group and the control group at T1: Practice sample.

Variable	Intervention Group ($n = 7$ from 5 schools)			Control Group ($n = 17$ from 8 schools)		
	n (%)[a]	*M*	*SD*	*n* (%)[a]	*M*	*SD*
Gender						
Female	7 (100)			13 (76.50)		
Male	-- (0)			4 (23.50)		
Federal state						
State 1	7 (100)			1 (5.90)		
State 2	-- (0)			16 (94.10)		
Mean age in years		40.33	4.04		42.69	8.16
Background in elementary school education	5 (71.40)			16 (94.10)		
Background in elementary school science	1 (14.30)			7 (41.20)		
Background in language support/German as a second language	4 (57.10)			3 (17.60)		

Note. The practice sample includes all teachers who participated in the project throughout the entire professional development phase (incl. written assignments of pretests and posttests), and also took part in the implementation phase.
[a] Information is based on the valid percentages.

elementary school teachers in the participating states and in Germany overall (OECD, 2020). Regarding their educational background, no substantial differences emerged for the number of teachers who had studied science as a school subject ($\chi^2(1) = 2.23$, $p = .14$, $\varphi = .27$, 95% CI [.00, .62]). However, there were more teachers who had studied elementary school education in the CG than in the IG ($\chi^2(1) = 3.13$, $p = .08$, $\varphi = .32$, 95% CI [.00, .67]). More IG-teachers than CG-teachers had taken courses on language support or German as a second language during their university education ($\chi^2(1) = 4.58$, $p = .03$, $\varphi = .40$, 95% CI [.00, .76]). Moreover, all IG-teachers were located in State 1, whereas the majority of the CG-participants taught at schools in State 2 ($\chi^2(1) = 24.69$, $p < .001$, $\varphi = .88$, 95% CI [.53, 1.22]).

3.2.1. Sample Attrition

Due to challenges such as teacher shortage during teacher recruitment and sample attrition during the PD phase, we find an uneven distribution of IG and CG across locations. The project had initially been planned to be conducted only in one state, yet sample attrition forced us to additionally recruit teachers in another state during the ongoing project. Considering the time-consuming PD for language support, teachers from State 1 were preferably assigned to the IG. Participants from State 2, who joined the project at a later time point, were allocated to the CG which involved a less time-consuming PD. Altogether, 12 teachers dropped out before completing the PD phase and could not be included in the present analyses, resulting in a total of 32 teachers in the cognition sample. Teachers who dropped out had less frequently studied elementary school education than those who remained in the project ($\chi^2(1) = 4.33$, $p = .04$, $\varphi = .32$, 95% CI [.00, .62]). No differences emerged in teachers' frequency of studying elementary school science ($\chi^2(1) = 0.43$, $p = .51$, $\varphi = .10$, 95% CI [.00, .40]) or attending courses on language support or German as a second language during their university education ($\chi^2(1) = 1.16$, $p = .28$, $\varphi = .17$, 95% CI [.00, .49]). During the implementation phase, the project was reduced by eight more teachers, leaving a total of 24 teachers in the practice sample. The dropouts did not differ from those who participated throughout the implementation phase regarding their background in elementary school education ($\chi^2(1) = 1.40$, $p = .24$, $\varphi = .21$, 95% CI [.00, .56]), elementary school science ($\chi^2(1) = 0.19$, $p = .89$, $\varphi = .02$, 95% CI [.00, .30]), and training on language support or German as a second language ($\chi^2(1) = 0.19$, $p = .67$, $\varphi = .08$, 95% CI [.00, .44]). Staff shortage and overstraining were among the most important reasons for the severe sample attrition during PD and implementation phase, along with various personal and organizational reasons (e.g., teacher's placement in another grade level with no opportunity to teach a Grade 3 science classroom; teacher's change of school or change of headmaster within a school with no project support by the new headmaster) that do not suggest a systematic dropout.

3.3. PD program for intervention group and control group

3.3.1. PD for elementary school science

To make sure that teachers in the IG and in the CG had a comparable knowledge base on the relevant elementary school science topics, teachers of both groups took part in PD courses on the topics of “floating and sinking” (Topic 1) and “evaporation and condensation” (Topic 2). Both topics are part of the elementary school science curricula of the participating federal states. The PD classes were conducted by two authors of the present paper with a background in elementary school science and didactics. Each course comprised 5 h and aimed at developing the CK and PCK necessary for teaching the respective lesson units. The curriculum on “floating and sinking” focused on the concepts of density, water displacement, pressure, and buoyancy force (Kleickmann et al., 2016; Möller et al., 2002); the curriculum on “evaporation and condensation” addressed the hydrological cycle and the processes of evaporation and condensation (for a detailed description, see Supplementary Materials A and B).

In line with the previously described characteristics of effective PD (e.g., Darling-Hammond et al., 2009; Lipowsky & Rzejak, 2015), both PDs had a clear content-specific focus, provided ample opportunity for collaboration among participants, and combined phases of input, practice, and reflection. Specifically, teachers received input regarding important physics principles and they reflected on typical student misconceptions and instructional approaches. Moreover, the lesson units were briefly introduced and teachers had the opportunity to work with the teaching aids (e.g., materials such as real objects in different sizes and from different materials, work sheets), to conduct experiments, and to discuss their observations. Upon course completion, teachers received all materials needed for teaching the lesson units in class. This included manuals for both topics with detailed lesson plans for each lesson.

3.3.2. PD for language support in science classrooms

The present study applied a newly developed PD program for language support in science classrooms. The language-support approach followed the scaffolding principles (Darsow et al., 2012; Gibbons, 2002) and aimed at promoting teachers' language-support skills in terms of both theoretical background knowledge and application to instruction (Gabler et al., 2020). Drawing on common characteristics of effective PD, we took great care to integrate phases of input with application and reflection. The PD included workshops and coaching and engaged participants in active collaboration (for a detailed description, see Supplementary Material C). One of the authors of the present paper, who is an expert in the field of linguistics and language support, primarily developed and implemented the PD program.

Module 1: Basics of language scaffolding. This part of the PD encompassed three training sessions (2 half-day sessions, 1 full-day session, 16 h altogether) in which the central components of the scaffolding approach were introduced and applied to the teaching unit of “floating and sinking”. The scaffolding approach includes an extensive lesson planning phase with a strong focus on the formulation of language-related learning goals (macro-scaffolding), and the actual language support in class (micro-scaffolding). For defining language-related learning goals, teachers need to be aware of a topic's linguistic requirements and they need to be able to evaluate their students' language skills. In Module 1 of the PD, the challenges of academic language in general and the subject-specific challenges of the topic “floating and sinking” in particular (e.g., linguistic knowledge needed for formulating a hypothesis) were introduced. For each lesson, the teachers received detailed descriptions of the linguistic demands students need to master in order to develop the underlying concepts, and they learned how to analyze the linguistic requirements of teaching materials (macro-scaffolding). Based on examples from publicly available lesson transcripts and videos (e.g., https://www.uni-muenster.de/Koviu/; cf. Steffensky et al., 2015), the participants identified situations that are most suitable for language support. Furthermore, teachers discussed formal (i.e., test instruments) and informal ways (e.g., students oral and written utterances in class) of assessing students' language skills and worked on the definition of language-related learning goals. To actively support students' language acquisition in classroom interaction (micro-scaffolding), teachers learned a variety of implicit and explicit language-support strategies (for a detailed description, see Table 1).

In the last training session, IG-teachers received an updated version of the manual for “floating and sinking” that additionally included didactical and methodological comments on language support. This updated version of the manual served as a wrap-up of the topics that had been covered in the previous training sessions and teachers were thoroughly familiarized with it (e.g., within role plays of selected classroom situations, by reflecting on the applicability of the suggested hints for language support for their own classroom teaching).

Module 2: Coaching and video feedback. In Module 2, which typically started 1–2 weeks after Module 1, IG-teachers delivered at least two lessons of the curriculum on “floating and sinking” in their regular Grade 3 or Grade 4 classrooms (these were different classrooms than those who took part in the implementation phase in the subsequent school year). Two of these lessons were followed by extensive coaching sessions with reflection and feedback. The coaching sessions took place in a one-on-one setting (teacher and trainer) or in a small group setting (4–6 teachers and trainer) and without the students' participation. The trainer observed the first lesson in the participating teachers’ classes. In the subsequent one-on-one feedback, teachers had the chance to reflect on the attainment of the language-support goals they had set for themselves, they received feedback on successful language support, were shown opportunities for further development, and formulated new goals for the next lesson.

We videotaped the second lesson and used selected video clips for video-feedback sessions in small groups. In these meetings, participants discussed both particularly successful sequences of language support and jointly developed ideas for further improvements. The video-feedback sessions took between 2 and 3 h and each IG-member participated in one such meeting. A final reflection meeting of all IG-members (3 h) was aimed at integrating the insights from the small group sessions.

3.4. Measures

3.4.1. Measures for evaluating PD effects on teachers’ cognition

Language-support skills across the curriculum. We used the LASSKI (LAnguage-support SKIlls scale), a researcher-developed measure for assessing teachers' language-support skills across the curriculum. The LASSKI focuses on the contents of the PD program and captures knowledge and application of the scaffolding components. It encompasses nine multiple choice (MC) or forced choice items (coded with 0 and 1) and two constructed response (CR) items (coded with 0, 1, and 2), resulting in a maximum score of 12 (see Supplementary Material D). Two trained student assistants who were blind to treatment conditions coded the CR items based on detailed scoring guidelines. Interrater reliability was moderate to very good (pretest: $.65 \leq \kappa \leq .83$, posttest: $.59 \leq \kappa \leq .83$; pretest: $M_\kappa = .74$, $SD_\kappa = .13$, posttest: $M_\kappa = .71$, $SD_\kappa = .17$). When the coders’ ratings differed, we used their mean for further analyses. The reliability of the scale was satisfactory in the pretest ($\alpha = .74$) but not sufficient in the posttest ($\alpha = .66$).

PCK on floating and sinking. Drawing on 5 open-ended questions that were included in the evaluation of a previous version of the PD (Decker et al., 2020), we developed an instrument for assessing teachers' PCK on “floating and sinking”. The revised instrument tapped the domains “knowledge on students’ understanding” (including typical preconcepts) and “knowledge on teaching strategies” (cf. Park & Oliver, 2007) and was closely aligned with the contents of the teacher PD on “floating and sinking”. For example, teachers were asked to provide typical student explanations why a boat floats on the water or deliver a sketch for visualizing the concept of density (see the Supplementary Material A for an overview of the PD contents). The instrument consisted of eight tasks (six CR and two MC) with 18 items. Answers were coded with 0, 1, and 2 and the maximum score was 24. Using a detailed coding manual, all items were independently rated by two trained student assistants in a blind coding procedure. Interrater reliability ranged from low to excellent (pretest: $.47 \leq \kappa \leq 1.00$, posttest: $.30 \leq \kappa \leq 1.00$) and was very good for the majority of items (pretest: $M_\kappa = .85$, $SD_\kappa = 017$, posttest: $M_\kappa = .78$, $SD_\kappa = .24$). In case of divergent ratings, the coders discussed these and agreed upon a rating, which was used for subsequent analyses. The reliability of the scale was satisfactory ($\alpha_{pre/post} = .79/.76$).

PCK on evaporation and condensation. For assessing teachers' PCK on “evaporation and condensation”, we used a revised version of the measure by Lange et al. (2012) which included the two domains “knowledge on students' understanding” and “knowledge on teaching strategies”. The instrument thus tapped teachers’ knowledge of frequent student preconcepts of evaporation, possible difficulties for understanding this transition process, and the purpose of typical classroom experiments (such as boiling water in a pot), for instance. These contents are generally covered by elementary school science curriculums on "evaporation and condensation" and were also part of our PD (see Supplementary Material B). The final scale had seven tasks with 18 items, 16 of which were CR items and two required participants to draw a sketch (of a water cycle, for instance). Answers were coded with 0, 1, or 2 points, resulting in a maximum score of 36. The coding and scoring procedure followed the same principle as for the PCK-test on "floating and sinking". Interrater reliability was moderate to very good for all but one item (pretest: $.23 \leq \kappa \leq 1.00$, posttest: $.40 \leq \kappa \leq 1.00$; pretest: $M_\kappa = .71$, $SD_\kappa = .20$, posttest: $M_\kappa = .76$, $SD_\kappa = .20$). Coders discussed divergent ratings and we used their consensus rating for further analyses. The reliability of the scale was not sufficient in the pretest ($\alpha = .64$) but satisfactory in the posttest ($\alpha = .78$).

Self-efficacy for teaching elementary school science. We used a researcher-developed scale to assess teachers' self-efficacy for teaching elementary school science. On this scale, teachers indicated to what degree they believed they could explain or demonstrate phenomena that are typical for the elementary school science topics "floating and sinking", "condensation and evaporation", or "education for sustainable development" (e.g., "How confident are you that you could convey the difference between an object's density and its weight to your students?"). All questions were answered on a 4-point Likert scale (1 = very unconfident, 4 = very confident). The scale consisted of 14 items and its reliability was very good ($\alpha_{pre/post} = .90/.86$).

3.4.2. Measures for evaluating PD effects on teachers' classroom practice

We used the Classroom Assessment Scoring System (CLASS K-3; Pianta et al., 2008) to assess teachers' instructional support during science classes. The CLASS is an observation instrument for measuring interaction quality in the classroom. In the present study, we focused on the domain of "instructional support", which includes the dimensions "concept development", "quality of feedback", and "language modeling". Concept development refers to activities aimed at promoting students' higher-order thinking skills and conceptual understanding (e.g., by activating prior knowledge or by engaging students in experiments). These also played an important role and were deliberately encouraged in the inquiry-based science curricula of the present study. Quality of feedback captures the amount of feedback a teacher gives that helps expand learning and understanding (e.g., by providing additional information and by encouraging students). Language modeling taps "the quality and amount of the teacher's use of language-stimulation and language-facilitation techniques" (Pianta et al., 2008, p. 79). The quality of each of these domains is assessed on a 7-point rating scale (1–2: low quality; 3–5: average quality; 6–7: high quality) and evaluations are based on a range of indicators. Language modeling, for instance, involves an appraisal of the use of advanced language (e.g., variety of words), the frequency of conversations, the use of open-ended questions, and self- and parallel talk (Pianta et al., 2008). There is substantial overlap between these indicators and the strategies included in the scaffolding approach (micro-scaffolding) that were taught in the PD for language support (see Table 1 and Supplementary Material C).

Given the highly-inferential nature of the CLASS, ratings must be carried out by trained and licensed raters. In the present study, external licensed raters who were blind to our study goals and treatment conditions carried out all CLASS ratings. The ratings were based on 20-min videoclips of the two videotaped double lessons from the teaching units on Topics 1 and 2. Each video (Topic 1: $n = 24$, Topic 2: $n = 24$) was coded by two raters and interrater-reliability was satisfactory to very good for both topics. The intraclass correlation coefficients were .80 (concept development), .63 (quality of feedback), and .75 (language modeling) for Topic 1, and .84 (concept development), .88 (quality of feedback), and .83 (language modeling) for Topic 2. In line with the CLASS guidelines, we did not change the individual ratings for our analyses but used mean scores when ratings differed.

3.5. Statistical analyses

To test the PD effects on the cognition measures and compare IG and CG in their classroom behavior, we conducted a series of univariate and multivariate analyses of (co)variance. Regarding the preconditions for these analyses, within-group homoscedasticity was given for all but one measure (pretest score of the LASSKI). Given the small sample size, some of the data were not normally distributed within IG and CG. Yet, ANOVAS are typically considered and have empirically been shown (Blanca et al., 2017; Schmider et al., 2010) to be relatively robust against violations of the normality assumption. As statistical significance depends on sample size, even meaningful effects might not be statistically significant in small samples. For evaluating the practical relevance of our findings, we therefore report the effect size partial η^2 for all results, with .01 indicating a small, .06 a medium, and .14 a large effect (Cohen, 1988). The effect size estimates are, however, subject to a high level of inaccuracy, as indicated by the breath of the 95% CIs. All analyses were also run with nonparametric tests; the results obtained with nonparametric tests closely correspond to the results reported below.

4. Results

4.1. PD effects on teachers' cognition

In a first step, we explored comparability across groups on teachers' pretest scores for all cognition variables (for descriptive statistics and intercorrelations, see Table 4). Univariate analyses revealed no differences between IG and CG in their language-support skills, $F(1,27) = .19$, $p = .66$, partial $\eta^2 = .01$, 95% CI [.00, .16], their PCK on "evaporation and condensation", $F(1,27) = 0.31$, $p = .58$, partial $\eta^2 = .01$, 95% CI [.00, .18], and their self-efficacy for teaching elementary school science, $F(1,29) = 0.87$, $p = .36$, partial $\eta^2 = .03$, 95% CI [.00, .21]. Pretest differences on teachers' PCK on "floating and sinking" did not yield statistical significance, $F(1,28) = 3.81$, $p = .06$, partial $\eta^2 = .12$, 95% CI [.00, .34], but the effect size indicated a medium effect, thus suggesting that the CG entered the PD with better PCK than the IG. This finding supports our decision to foster teachers' knowledge on the relevant elementary school science topics as a precondition for investigating the effectiveness of the PD on language support.

In line with our study aims, we next examined PD effects for knowledge on language support. As is typically suggested for experimental designs with a pretest and a posttest (e.g., Rausch et al., 2010), we controlled for pretest scores when comparing IG and CG in their posttest performance on the LASSKI (i.e., our measure of teachers' language-support skills). This increases the power of analyses compared to a repeated-measures ANOVA (Table 5). The ANCOVA revealed a significant group difference with a large effect size, $F(1,21) = 6.33$, $p = .02$, partial $\eta^2 = .23$, 95% CI [.00, .48]. Results indicate that the IG outperformed the CG, hence supporting the effectiveness of the PD for improving teachers' language-support skills.

The PD for elementary school science did not include an experimental factor but was the same for IG and CG. To test

Table 4
Descriptive statistics and correlations for the cognition measures.

#	Time Point	Measure	Intervention Group			Control Group			1	2	3	4	5	6	7
			n	*M*	*SD*	*n*	*M*	*SD*							
1	Pretest	Language-support skills	10	5.20	1.38	19	4.74	3.16	–						
2	Posttest	Language-support skills	9	8.61	2.45	18	6.86	2.17	.19	–					
3	Pretest	PCK on floating and sinking	10	6.80	4.59	20	10.00	4.05	-.03	-.13	–				
4	Posttest	PCK on floating and sinking	10	11.10	3.93	18	11.17	3.75	.01	.16	.51**	–			
5	Pretest	PCK on evaporation and condensation	10	12.90	4.98	19	13.84	3.93	.05	.09	.30	.26	–		
6	Posttest	PCK on evaporation and condensation	6	15.67	6.31	13	14.38	5.91	-.10	.22	.27	.39	.00	–	
7	Pretest	Self-efficacy for teaching elementary school science	11	2.67	0.53	20	2.84	0.49	-.09	-.21	.41*	-.06	.35+	.37	–
8	Posttest	Self-efficacy for teaching elementary school science	10	3.04	0.56	20	2.87	0.43	-.24	.29	.04	.23	.09	.38	.36+

Note. + $p < .10$, *$p < .05$, **$p < .01$.

Table 5
Results of the repeated-measures ANOVAs for investigating PD effects.

Measure	Factor Time (Pretest, Posttest)				Factor Treatment Condition (IG, CG)				Interaction Time x Treatment Condition			
	F	*p*	partial η^2	95% CI	*F*	*p*	partial η^2	95% CI	*F*	*p*	partial η^2	95% CI
Language-support skills[a]	$F(1,22) = 21.01$	<.01	.49	[.03, .67]	$F(1,22) = 1.60$	.22	.07	[.00, .30]	$F(1,22) = 2.98$	.10	.12	[.00, .36]
PCK on floating and sinking	$F(1,24) = 9.56$	.01	.29	[.03, .51]	$F(1,24) = 1.74$	.20	.07	[.00, .29]	$F(1,24) = 2.71$	.11	.10	[.00, .34]
PCK on evaporation and condensation	$F(1,16) = 0.89$	.36	.05	[.00, .32]	$F(1,16) = 0.05$	.82	<.01	[.00, .18]	$F(1,16) = 0.26$	.62	.02	[.00, .25]
Self-efficacy for teaching elementary school science	$F(1,27) = 4.72$	.04	.15	[.00, .38]	$F(1,27) = 0.02$	.90	<.01	[.00, .09]	$F(1,27) = 3.70$	.07	.12	[.00, .35]

Note. IG = intervention group, CG = control group. CI = confidence interval. The 95% CI refers to the effect size partial η^2.
[a] As we implemented an experimental design for the PD for language-support skills, we report the results of an ANCOVA, controlling for the pretest measures, in the body of the paper. This procedure increases the statistical power of the analyses compared to a repeated-measures ANOVA (Rausch et al., 2010). As a robustness check, we additionally performed a repeated-measures ANOVA, which we report here. Given the lower statistical power, this analysis revealed statistically insignificant but medium-sized effects for the treatment condition and the interaction time x treatment condition.

treatment effects, we conducted separate repeated-measures ANOVAs, modelling time (pretest, posttest) as within-subject factor and treatment condition (IG, CG) as between-subject factor. Teachers' PCK on Topics 1 and 2, as well as their self-efficacy for teaching elementary school science served as dependent variables. A very similar pattern of results emerged for PCK on Topic 1 and for teachers' self-efficacy for teaching elementary school science (Table 5). In both cases, we found a significant main effect for time, indicating that participants in both groups increased their PCK as well as their self-efficacy over time. A significant group effect did not emerge. Yet, we found a medium-sized, albeit not significant interaction for group x time for both measures. These suggest that the development from pretest to posttest may have been somewhat stronger for the IG than for the CG.

For teachers' PCK on Topic 2, however, no significant main or interaction effects were found (Table 5). This means that the PD on “evaporation and condensation” did not increase teachers' PCK, as measured by our test, neither for the IG nor for the CG.

4.2. PD effects on teachers' classroom practice

We used the practice sample (Table 3) for investigating PD effects on teachers‘ classroom practice. Before investigating potential IG and CG-differences in teachers' language-supportive teaching, we checked whether participants fully implemented the lesson units on Topics 1 and 2. We used the videotaped lessons (90 min) of both science topics and examined whether compulsory elements of the lesson plans occurred in the lessons. On average, teachers implemented 90.03% ($SD = 10.26\%$) of the elements included in the lesson plan for Topic 1 and 88.97% ($SD = 9.55\%$) of the elements of the lesson plan for Topic 2. While the IG tended to implement fewer elements of the lesson plans of both topics (Topic 1: $t = 1.08$, $df = 21$, $p = .29$, $d = 0.49$, 95% CI [-0.42, 1.38]; Topic 2: $t = 1.37$, $df = 22$, $p = .18$, $d = 0.62$, 95% CI [-0.29, 1.51]), both groups showed a high degree of implementation fidelity.

Table 6 displays the descriptive statistics and correlations for the classroom instruction variables. CLASS scores from 3 to 5 reflect average instructional quality, thus indicating that participants of both groups obtained medium results in all measurements except for the CG results of quality of feedback for Topic 1. Post-hoc tests revealed that both groups improved their instructional support from Topic 1 to Topic 2 (Table 7). Moreover, the three dimensions of instructional quality are strongly correlated within each teaching unit (Table 6), while correlations between the same dimensions across topics are smaller. This suggests a relatively low stability of instructional quality across teaching units.

To answer our research question and test potential differences between IG and CG in concept development, quality of feedback, and language modeling, we performed two MANOVAs, one for Topic 1 and one for Topic 2. We entered the treatment condition as independent variable and the three CLASS dimensions of instructional support as dependent variables. For Topic 1, the IG was rated significantly higher on quality of feedback, resulting in a large effect size, $F(1,22) = 4.96$, $p = .04$, partial $\eta^2 = .18$, 95% CI [.00, .43]. As expected, the IG also outperformed the CG on language modeling, yielding a medium albeit not significant effect, $F(1,22) = 2.10$, $p = .16$, partial $\eta^2 = .09$, 95% CI [.00, .33]. No group differences emerged for concept development, $F(1,22) < 0.01$, $p = .98$, partial $\eta^2 < .01$, 95% CI [.00, .00]. Contrary to our expectations, IG and CG did not significantly differ on any of the three dimensions of

Table 6
Descriptive statistics and correlations for the practice measures by science topic.

#	Science Topic	Measure	Intervention Group			Control Group			1	2	3	4	5
			n	*M*	*SD*	*n*	*M*	*SD*					
1	Floating and sinking (Topic 1)	Concept development	7	2.93	0.98	17	2.94	0.95	–				
2		Quality of feedback	7	3.00	0.65	17	2.41	0.57	.45*	–			
3		Language modeling	7	3.79	1.19	17	3.21	0.75	.70***	.56**	–		
4	Evaporation and condensation (Topic 2)	Concept development	7	4.00	0.91	17	4.47	0.87	-.07	.05	-.17	–	
5		Quality of feedback	7	3.14	0.75	17	3.53	1.12	.39	.33	.16	.59**	–
6		Language modeling	7	4.14	0.80	17	4.00	0.94	.04	.48*	.24	.63**	.52**

Note. *$p < .05$, **$p < .01$, ***$p < .001$.

Table 7
Results of the post-hoc tests for comparing the practice variables across science topics.

Measure	Factor Time				Factor Treatment Condition			
	(Pretest, Posttest)				(IG, CG)			
	F	*p*	partial η^2	95% CI	*F*	*p*	partial η^2	95% CI
Concept development	$F(1,22) = 18.47$	<.01	.46	[.13, .64]	$F(1,22) = 0.73$	.40	.03	[.00, .15]
Quality of feedback	$F(1,22) = 9.06$	.01	.29	[.03, .52]	$F(1,22) = 0.10$	.75	.01	[.00, .03]
Language modeling	$F(1,22) = 5.28$	.03	.19	[.00, .44]	$F(1,22) = 1.32$	.26	.06	[.00, .29]

Note. IG = intervention group, CG = control group. CI = confidence interval. The 95% CI refers to the effect size partial η^2.

instructional support in the lesson on Topic 2, but the CG tended to show a slight advantage on concept development (concept development: $F(1,22) = 1.40$, $p = .25$, partial $\eta^2 = .06$, 95% CI [.00, .29]; quality of feedback: $F(1,22) = 0.69$, $p = .41$, partial $\eta^2 = .03$, 95% CI [.00, .25]; language modeling: $F(1,22) = 0.12$, $p = .73$, partial $\eta^2 = .01$, 95% CI [.00, .17]). Group differences were thus not stable across time points.

5. Discussion

The present study investigated the effects of a newly developed PD program aimed at improving in-service teachers' skills for providing language support in regular elementary school science classrooms. The PD program was implemented over the course of a full school year. It included PD courses on important elementary school science topics (for IG and CG) as well as theoretical and application-oriented training on content-integrated language support based on the scaffolding approach (for IG only). We used outcome measures of teacher cognitions and classroom practices to evaluate the program, hence focusing on evaluation Levels 2 and 3 (Kalinowski et al., 2020; Lipowsky & Rzejak, 2015).

Regarding teachers' knowledge on language support, we found that the IG outperformed the CG in the posttest, controlling for pretest scores, thus pointing to the effectiveness of the intervention. As for the PD on elementary school science, results revealed positive treatment effects on teachers' PCK on Topic 1 ("floating and sinking") and on their self-efficacy for teaching elementary school science, but not on their PCK on Topic 2 ("condensation and evaporation").

With regard to teachers' classroom practice, we expected the IG to outperform the CG on the dimension of language modeling, but not on the dimensions of concept development and quality of feedback, as the latter two do not specifically focus on language-stimulation and language-facilitation (Pianta et al., 2008). Respective differences between IG and CG were observed for Topic 1. Specifically, quality of feedback was better in the IG than in the CG. IG-teachers also tended to incorporate language modeling to a higher degree into their lesson on Topic 1 than CG-teachers. For Topic 2, no significant effects emerged. Although group differences for Topic 1 were not sustained over time, we found that both groups provided better instructional quality in Topic 2 than in Topic 1, as reflected in all three CLASS dimensions.

Previous evaluation studies, most of which were conducted in the US and focused on ELLs, provided evidence that teachers may profit from PD for language support in various respects (e.g., self-efficacy for forstering students language skills during content classes, self-assessed knowledge; Cantrell & Hughes, 2008; Hart & Lee, 2003; Kalinowski et al., 2020; Kalinowski et al., 2019; van Dijk et al., 2019). This is basically in line with our finding that an intensive teacher PD on content-integrated language support, combining phases of input, practice, and reflection (cf. Lipowsky & Rzejak, 2015), contributes to teachers' knowledge and, at least to a certain extent, to classroom practice. Yet, whereas we identified larger PD effects on teachers' cognition (measured by knowledge on content-integrated language support) than on language-supportive classroom practices, Kalinowski et al. (2020) report the reverse pattern in their meta-analysis.

In interpreting the results, it should be considered that research on the effectiveness of teacher PD for (academic) language support in mainstream classrooms is still at an early stage and that there is a large variability across studies in terms of underlying PD concepts, study designs, and methods. The number of studies focusing on treatment effects at Level 2 is generally small and most published studies relied on self-assessments in measuring effects on teachers' knowledge and beliefs (e.g., Cantrell & Hughes, 2008; Lee & Maerten-Rivera, 2012). Our results, in contrast, are based on data collected with an objective test instrument. Furthermore, studies investigating PD effects on teachers' language-support behavior in the classroom typically employ newly developed instruments that are closely aligned with the PD contents but have not been used and validated independently of the specific study (for an overview, see Kalinowski et al., 2020). Considering that treatment effects tend to be larger on researcher-developed measures than on well-established, standardized instruments (Babinski et al., 2018; Kalinowski et al., 2020), the smaller IG-CG-differences in the present study may, at least in part, be due to characteristics of our observation instrument. The well-established CLASS instrument includes various important aspects of language-supportive teaching that were focused in our PD (e.g., self- and parallel talk, open-ended questions) but does not specifically draw on the scaffolding approach for fostering students' language development (Pianta et al., 2008).

Furthermore, a core characteristic of the PD program may have contributed to the IG-CG-differences that emerged for Topic 1 but not for Topic 2, i.e., that IG teachers showed larger learning gains in PCK for Topic 1 and that IG and CG differed in language modeling and quality of feedback during the lesson on Topic 1, but not on Topic 2. As we took great care to help teachers integrate language-support strategies into their regular science teaching, the PD program for language support was embedded in the framework of the lesson unit on “floating and sinking”. This means that the scaffolding approach was introduced using examples from the teaching unit on Topic 1, the lesson plans contained suggestions for language support, and teachers implemented at least two lessons of the curriculum for Topic 1 during the application phase of the PD. IG-teachers thus had substantially more opportunities than CG-teachers to familiarize themselves with the “floating and sinking” curriculum and could benefit from specific suggestions for integrating language-support strategies into their subject-area teaching. In contrast, both groups received identical PD and teaching materials (without didactical comments on language support) for Topic 2.

Results further indicated that teachers of both groups showed a higher degree of instructional support in teaching Topic 2 than in teaching Topic 1. This might be due to specific characteristics of the curriculum on “floating and sinking”, which included challenging concepts, such as water displacement and buoyancy force. It also involved the targeted use of a multitude of teaching materials (e.g., water basins and cups in different sizes, real objects such as dices and balls from different materials) and was therefore highly demanding with regard to classroom organization.

The amount of opportunities to learn may also help explain the finding that the PD on Topic 2 did not benefit teachers' PCK on “evaporation and condensation”. For this topic, both groups participated in a 5-h PD course but did not teach the contents in their classes before taking the written posttests. Considering the IG's learning gains on Topic 1 and their intense engagement with this topic over a longer period of time, the null effects for Topic 2 might indicate that the theoretical training needs to be extended by practical teaching experience in order to effectively increase teachers' PCK. Although the PD training combined phases of input, reflection, and practice (e.g., Darling-Hammond et al., 2009), the dosage of a single PD course probably was not enough for yielding measurable effects.

5.1. Limitations

There are a number of limitations to the present investigation. First and foremost, analyses are based on a small convenience sample of 32 in-service teachers (24 for the practice variables) from two German federal states. Sample attrition during the PD phase resulted in a confounding of treatment condition and location. Whereas the IG reported more opportunities to learn on the topic of language support during their university education than the CG, thus suggesting greater prior experience with and interest in the topic, the groups did not differ in their pretest scores on the measure for language-support skills. Pretest differences between IG and CG only occurred for PCK on Topic 1 and were compensated by the PD. While the effects of the confounding thus seem to be negligible, the results still need to be interpreted with caution and cannot be generalized to elementary school teachers overall. Moreover, due to the small sample size, the estimation of the effect sizes is subject to a relatively high degree of inaccuracy as reflected in the breadth of the confidence intervals. These indicate that, in many instances, both relatively strong effects but also null effects might be detected when repeating the study. While acknowledging this shortcoming, it is worth mentioning that small sample sizes are a common, and perhaps even inevitable, challenge of studies aimed at evaluating complex, intensive, and time-consuming teacher PD for language support. Of the studies included in the meta-analysis by Kalinowski et al. (2020), for instance, six out of 10 included less than 35 participants (for further examples not included in the meta-analysis, see Batt, 2010; Rivard & Gueye, 2016; Tong et al., 2018; van Dijk et al., 2019).

Second, we neither interviewed teachers regarding their teaching practices, nor videotaped their classroom teaching before participating in the PD program. Pre-post-comparisons were therefore only possible for the written assignments but not for teachers' instructional support. Although it seems reasonable that differences between IG and CG may be due to the different treatment conditions, given the PD effects on teachers' language-support skills, it is not possible to test this directly with our data. We did, however, establish a high degree of comparability within and across groups by having all teachers implement the same curricula and by videotaping the exact same lessons in every classroom. Therefore, factors that could bias the assessment of teachers' instructional quality (e.g., different work phases with different opportunities for language modeling, topics that are linguistically more challenging than others) were minimized.

Third, although we analyzed both cognition measures and practice measures, our study included only a small range of outcome variables. Specifically, the present study does not allow for investigating teachers' implementation of the macro-scaffolding components of the scaffolding approach (i.e., analysis of a topic's linguistic demands, formulation of language-related learning goals during lesson planning, etc.). These preparatory steps are particularly important for targeted language support but tend to be implemented very infrequently (Elstrodt-Wefing et al., 2019; Vock et al., 2020). As a result, the language-support strategies used in classroom teaching may not be as efficient as they could be if they referred more deliberately to the specific linguistic demands of a given topic.

6. Conclusion

The present findings support the notion that teacher PD, which is delivered over an extended time period and engages participants in active learning phases, combined with feedback and reflection, contributes substantially to teachers' knowledge for subject-integrated language support and may support them in incorporating language-support strategies into their regular science teaching. Given the limited number of studies that systematically evaluated effects of teacher trainings in language support across the curriculum, particularly outside the US, results of the present investigation may motivate future research and practice. Having established measurable effects at Levels 2 and 3, further research is needed to examine whether teachers' learning gains translate into students' (academic) language development and, if so, which aspects of teachers' instructional behavior are particularly beneficial.

Taking a broader perspective on teacher training for subject-integrated language support, effective PD for in-service teachers is needed in different countries (e.g., Lucas et al., 2008), including Germany (Paetsch & Heppt, in press). The present PD yielded promising results, increasing participants' language-support skills and behavior. The curricula employed are widely applicable to early science instruction based on core science concepts. Given the usefulness of the scaffolding approach with respect to other subjects and grade levels, the present PD may lay the foundation for further PD, including programs that are aimed at secondary school teachers. Yet, our PD also entails characteristics that may impede its implementation on a broader scale (cf. Paetsch & Heppt, in press). It comprised several time and labor-intensive training and coaching sessions and required teachers' ongoing participation for an entire school year, although most teacher PDs in Germany are substantially shorter (Morris-Lange et al., 2016; Richter et al., 2020). Considering that teachers in Germany typically report a high work load, thus limiting their time for participating in PD (Richter et al., 2020), the individual commitment needed to participate in PD over such an extended period of time constitutes a challenge for both implementation and evaluation (cf. Geldenhuys & Oosthuizen, 2015). Moreover, the relatively high level of standardization of our intervention study is not easily achieved in practice. Therefore, implementation research is needed on the feasibility and effectiveness of possible program adaptations that can help prepare teachers for language-supportive teaching across the curriculum.

CRediT Author Statement

Birgit Heppt: Conceptualization, Investigation, Formal analysis, Writing – Original draft. Sofie Henschel: Conceptualization, Funding acquisition, Supervision, Writing – Review & Editing. Ilonca Hardy: Conceptualization, Funding acquisition, Supervision, Writing – Review & Editing. Rosa Hettmannsperger-Lippolt: Investigation, Project administration, Data curation. Katrin Gabler: Investigation, Project administration. Christine Sontag: Investigation, Project administration, Data curation. Susanne Mannel: Investigation. Petra Stanat: Writing – Review & Editing, Funding acquisition.

Acknowledgements

The CLASS ratings were conducted by Kathrin Kohake, Jessica Maier, and Alfred Richartz, Universität Hamburg, Germany. We sincerely thank them for their support.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) through Grant 01JI1602A, awarded to Humboldt-Universität zu Berlin, Germany and through Grant 01JI1602B awarded to Goethe University Frankfurt, Germany. The authors assume full responsibility of the content of the present publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2021.103518.

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H.-L. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397–419. https://doi.org/10.1002/sce.10118

Armour, K. M., & Yelling, M. (2007). Effective professional development for physical education teachers: The role of informal, collaborative learning. *Journal of Teaching in Physical Education, 26*(2), 177–200. https://doi.org/10.1123/jtpe.26.2.177

August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D. J., Powell, J., Moore, S., & Haynes, E. F. (2014). Helping ELLs meet the common core state standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness, 7*(1), 54–82. https://doi.org/10.1080/19345747.2013.836763

Babinski, L. M., Amendum, S. J., Knotek, S. E., Sánchez, M., & Malone, P. (2018). Improving young English learners' language and literacy skills through teacher professional development: A randomized controlled trial. *American Educational Research Journal, 55*(1), 117–143. https://doi.org/10.3102/0002831217732335

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.

Batt, E. G. (2010). Cognitive coaching: A critical phase in professional development to implement sheltered instruction. *Teaching and Teacher Education, 26*(4), 997–1005. https://doi.org/10.1016/j.tate.2009.10.042

Becker-Mrotzek, M., & Roth, H.-J. (2017). Sprachliche Bildung – Grundlegende Begriffe und Konzepte [Language support across the curriculum – basic terms and concepts]. In M. Becker-Mrotzek, & H.-J. Roth (Eds.), *Sprachliche Bildung – Grundlagen und Handlungsfelder* (pp. 11–36). Waxmann.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data. Is ANOVA still a valid option? *Psicothema, 29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Blomberg, G., Gamoran Sherin, M., Renkl, A., Glogger, I., & Seidel, T. (2014). Understanding video as a tool for teacher education: Investigating instructional strategies to promote reflection. *Instructional Science, 42*, 443–463. https://doi.org/10.1007/s11251-013-9281-6

Brisk, M. E., & Zisselsberger, M. (2011). We've let them in on the secret": Using SFL theory to improve the teching of writing to bilingual learners. In T. Lucas (Ed.), *Teacher preparation for linguistically diverse classrooms. A resource for teacher educators* (pp. 111–126). Taylor & Francis.

Bunch, G. C. (2013). Pedagogical language knowledge: Preparing mainstream teachers for English learners in the new standards era. *Review of Research in Education, 37*(1), 298–341. https://doi.org/10.3102/0091732X12461772

Cantrell, S. C., & Hughes, H. K. (2008). Teacher efficacy and content literacy implementation: An exploration of the effects of extended professional development with coaching. *Journal of Literacy Research, 40*(1), 95–127. https://doi.org/10.1080/10862960802070442

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Dannenbauer, F. M. (2002). Grammatik [grammar]. In S. Baumgartner (Ed.), *Sprachtherapie mit Kindern. Grundlagen und Verfahren* (pp. 105–161). Reinhardt.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession*. https://edpolicy.stanford.edu/sites/default/files/publications/professional-learning-learning-profession-status-report-teacher-development-us-and-abroad.pdf.

Darsow, A., Paetsch, J., & Felbrich, A. (2012). Konzeption und Umsetzung der fachbezogenen Sprachförderung im BeFo-Projekt [Conception and implementation of subject-related language support in the BeFo project]. In S. Jeuk, & J. Schäfer (Eds.), *Deutsch als Zweitsprache in Kindertageseinrichtungen und Schulen* (pp. 215–234). Fillibach.

Dawes, L. (2004). Talk and learning in classroom science. *International Journal of Science Education, 26*(6), 677–695. https://doi.org/10.1080/0950069032000097424

Decker, A.-T., Hardy, I., Hertel, S., Lühken, A., & Kunter, M. (2020). Die Entwicklung von fachdidaktischem Wissen und Überzeugungen von Grundschullehrkräften im Kontext von Lehrerfortbildungen zum „Schwimmen und Sinken“ [The development of primary school teachers' pedagogical content knowledge of the topic “floating and sinking” and beliefs about science teaching and learning in a professional development workshop]. *Psychologie in Erziehung und Unterricht, 67*(1), 61–76. https://doi.org/10.2378/peu2020.art06d

van Dijk, M., Menninga, A., Steenbeek, H., & van Geert, P. (2019). Improving language use in early elementary science lessons by using a video feedback intervention for teachers. *Educational Research and Evaluation, 25*(5–6), 299–322. https://doi.org/10.1080/13803611.2020.1734472

Elstrodt-Wefing, N., Starke, A., Möhring, M., & Ritterfeld, U. (2019). Umsetzung unterrichtsintegrierter Sprachförderung im Primarbereich. Eine Mixed-Methods-Untersuchung bei Lehrkräften in BiSS-Verbünden [Implementation of language support in primary education. A mixed-method study of teachers in BiSS associations]. *Empirische Sonderpädagogik, 11*(3), 191–209.

Fang, Z., & Wei, Y. (2010). Improving middle school students' science literacy through reading infusion. *The Journal of Educational Research, 103*(4), 262–273. https://doi.org/10.1080/00220670903383051

Fauth, B., Decristan, J., Decker, A., Büttner, G., Klieme, E., Hardy, I., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education, 86*, 102882. https://doi.org/10.1016/j.tate.2019.102882

Fukkink, R. G., Trienekens, N., & Kramer, L. J. C. (2011). Video feedback in education and training: Putting learning in the picture. *Educational Psychology Review, 23*(1), 45–63. https://doi.org/10.1007/s10648-010-9144-5

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Review of Educational Research, 82*(3), 300–329. https://doi.org/10.3102/0034654312457206

Gabler, K., Mannel, S., Hardy, I., Henschel, S., Heppt, B., Hettmannsperger-Lippolt, R., Sontag, C., & Stanat, P. (2020). Fachintegrierte Sprachförderung im Sachunterricht der Grundschule: Entwicklung, Erprobung und Evaluation eines Fortbildungskonzepts auf der Grundlage des Scaffolding-Ansatzes [Subject-integrated language support in elementary school subject classes: Development, testing and evaluation of a training concept based on the scaffolding approach]. In C. Titz, S. Weber, H. Wagner, A. Ropeter, C. Geyer, & M. Hasselhorn (Eds.), *Sprach- und Schriftsprachförderung wirksam gestalten: Innovative Konzepte und Forschungsimpulse* (pp. 59–83). Kohlhammer.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. https://

doi.org/10.3102/00028312038004915
Geldenhuys, J. L., & Oosthuizen, L. C. (2015). Challenges influencing teachers' involvement in continuous professional development: A South African perspective. *Teaching and Teacher Education, 51*, 203–212. https://doi.org/10.1016/j.tate.2015.06.010
Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*(3), 183–209. https://doi.org/10.1016/0010-0277(86)90034-X
Gentner, D. (2016). Language as cognitive tool kit: How language supports relational thought. *American Psychologist, 71*(8), 650–657. https://doi.org/10.1037/amp0000082
Gibbons, P. (2002). *Scaffolding language, scaffolding learning. Teaching second language learners in the mainstream classroom*. Heinemann.
Guskey, T. R. (2000). *Evaluating professional development*. Corwin.
Guskey, T. R., & Yoon, K. S. (2009). What works in professional development? *The Phi Delta Kappa, 90*(7), 495–500. https://doi.org/10.1177/003172170909000709
Hardy, I., Kloetzer, B., Moeller, K., & Sodian, B. (2010). The analysis of classroom discourse: Elementary school science curricula advancing reasoning with evidence. *Educational Assessment, 15*(3/4), 197–221. https://doi.org/10.1080/10627197.2010.530556
Hardy, I., Saalbach, H., Leuchter, M., & Schalk, L. (2020). Preschoolers' induction of the concept of material kind to make predictions: The effects of comparison and linguistic labels. *Frontiers in Psychology, 11*, 531503. https://doi.org/10.3389/fpsyg.2020.531503
Hart, J. E., & Lee, O. (2003). Teacher professional development to improve the science and literacy achievement of English language learners. *Bilingual Research Journal, 27*(3), 475–501. https://doi.org/10.1080/15235882.2003.10162604
Henrichs, L. F., & Leseman, P. P. M. (2014). Early science instruction and academic language development can go hand in hand. The promising effects of a low-intensity teacher-focused intervention. *International Journal of Science Education, 36*(17), 2978–2995. https://doi.org/10.1080/09500693.2014.948944
Heppt, B., Henschel, S., & Haag, N. (2016). Everyday and academic language proficiency: Investigating their relationships with school success and challenges for language minority learners. *Learning and Individual Differences, 47*, 244–251. https://doi.org/10.1016/j.lindif.2016.01.004
Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes & efficacy. *Education Policy Analysis Archives, 13*(10), 1. http://epaa.asu.edu/epaa/v13n10/.
Kalinowski, E., Egert, F., Gronostaj, A., & Vock, M. (2020). Professional development on fostering students' academic language proficiency across the curriculum—A meta-analysis of its impact on teachers' cognition and teaching practices. *Teaching and Teacher Education, 88*, 4–15. https://doi.org/10.1016/j.tate.2019.102971
Kalinowski, E., Gronostaj, A., & Vock, M. (2019). Effective professional development for teachers to foster students' academic language proficiency across the curriculum: A systematic review. *AERA Open, 5*(1), 1–23. https://doi.org/10.1177/2332858419828691
Klassen, R., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76. https://doi.org/10.1016/j.edurev.2014.06.001
Kleickmann, T., Tröbst, S., Jonen, A., Vehmeyer, J., & Möller, K. (2016). The effects of expert scaffolding in elementary science professional development on teachers' beliefs and motivations, instructional practices, and student achievement. *Journal of Educational Psychology, 108*(1), 21–42. https://doi.org/10.1037/edu0000041
Kunter, M., Kleickmann, T., Klusmann, U., & Richter, D. (2013). The development of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project* (pp. 63–77). Springer. https://doi.org/10.1007/978-1-4614-5149-5_4.
Lange, K., Kleickmann, T., Tröbst, S., & Möller, K. (2012). Fachdidaktisches Wissen von Lehrkräften und multiple Ziele im naturwissenschaftlichen Sachunterricht [Subject-related didactic knowledge of teachers and multiple objectives in lessons on natural sciences]. *Zeitschrift für Erziehungswissenschaft, 15*(1), 55–75. https://doi.org/10.1007/s11618-012-0258-z
Lee, O., & Maerten-Rivera, J. (2012). Teacher change in elementary science instruction with English language learners: Results of a multiyear professional development intervention across multiple grades. *Teachers College Record, 114*(8), 1–44.
Lipowsky, F., & Rzejak, D. (2015). Key features of effective professional development programmes for teachers. *RICERCAZIONE, 7*(2), 27–51.
Lucas, T., Villegas, A. M., & Freedson-Gonzalez, M. (2008). Linguistically responsive teacher education: Preparing classroom teachers to teach English language learners. *Journal of Teacher Education, 59*(4), 361–373. https://doi.org/10.1177/0022487108322110
Lucero, A. (2014). Teachers' use of linguistic scaffolding to support the academic language development of first-grade emergent bilingual students. *Journal of Early Childhood Literacy, 14*(4), 534–561. https://doi.org/10.1177/1468798413512848
Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32*(2), 265–302. https://doi.org/10.1017/S0272263109990520
Mahan, K. R. (2020). The comprehending teacher: Scaffolding in content and language integrated learning (CLIL). *The Language Learning Journal*. https://doi.org/10.1080/09571736.2019.1705879
Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction — What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496. https://doi.org/10.1002/tea.20347
Möller, K., Jonen, A., Hardy, I., & Stern, E. (2002). Die Förderung von naturwissenschaftlichem Verständnis bei Grundschulkindern durch Strukturierung der Lernumgebung [Promoting scientific understanding of primary school children by structuring the learning environment]. *Zeitschrift für Pädagogik - Beiheft, 45*, 176–191.
Morris-Lange, S., Wagner, K., & Altinay, L. (2016). *Lehrerbildung in der Einwanderungsgesellschaft: Qualifizierung für den Normalfall Vielfalt. [Teacher education in the immigration society: Qualification for the usual case of diversity]. Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR) GmbH*. https://www.svr-migration.de/wp-content/uploads/2017/05/SVR_FB_Lehrerbildung.pdf.
Mulholland, J., Dorman, J. P., & Odgers, B. (2004). Assessment of science teaching efficacy of preservice teachers in an Australian university. *Journal of Science Teacher Education, 15*(4), 313–331. https://doi.org/10.1023/B:JSTE.0000048334.44537.86
Ødegaard, M., Haug, B., Mork, S. M., & Sørvik, G. O. (2014). Challenges and support when teaching science through an integrated inquiry and literacy approach. *International Journal of Science Education, 36*(18), 2997–3020. https://doi.org/10.1080/09500693.2014.942719
OECD. (2020). Who are the teachers? In *Education at a glance 2020: OECD indicators*. OECD Publishing. https://doi.org/10.1787/27f5f9c5-en.
Oyoo, S. O. (2011). Language in science classrooms: An analysis of physics teachers' use of and beliefs about language. *Research in Science Education, 42*(5), 849–873. https://doi.org/10.1007/s11165-011-9228-3
Paetsch, J., & Heppt, B. (in press). Teaching German as a second language: Implications for teacher education in Germany. In J. Chin-Kin Lee & T. Ehmke (Eds.), Quality in teacher education and professional development: Issues and prospects in Chinese and German systems (pp. 253-267). Routledge.
Paetsch, J., Wolf, K. M., Stanat, P., & Darsow, A. (2014, 2014/03/01). Sprachförderung von Kindern und Jugendlichen aus Zuwandererfamilien [Language promotion of children and adolescents from immigrant families]. *Zeitschrift für Erziehungswissenschaft, 17*(2), 315–347. https://doi.org/10.1007/s11618-013-0474-1
Park, S., & Oliver, J. S. (2007). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*(3), 261–284. https://doi.org/10.1007/s11165-007-9049-6
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system: Manual K-3*. Paul H Brookes.
Piwowar, V., Thiel, F., & Ophardt, D. (2013). Training inservice teachers' competencies in classroom management. A quasi-experimental study with teachers of secondary schools. *Teaching and Teacher Education, 30*, 1–12. https://doi.org/10.1016/j.tate.2012.09.007
Prediger, S., & Zindel, C. (2017). School academic language demands for understanding functional relationships: A design research project on the role of language in reading and learning. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(7b), 4157–4188. https://doi.org/10.12973/eurasia.2017.00804a
Quílez, J. (2021). Supporting Spanish 11th grade students to make scientific writing when learning chemistry in English: The case of logical connectives. *International Journal of Science Education, 43*(9), 1–24. https://doi.org/10.1080/09500693.2021.1918794
Rausch, J. R., Maxwell, S. E., & Kelley, K. (2010). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology, 32*(3), 467–486. https://doi.org/10.1207/s15374424jccp3203_15
Richter, E., Marx, A., Huang, Y., & Richter, D. (2020). Zeiten zum beruflichen Lernen: Eine empirische Untersuchung zum Zeitpunkt und der Dauer von Fortbildungsangeboten für Lehrkräfte [Professional learning times: An empirical study on the timing and duration of in-service trainng courses for teachers]. *Zeitschrift für Erziehungswissenschaft, 23*(1), 145–173. https://doi.org/10.1007/s11618-019-00924-x
Rivard, L. P., & Gueye, N. R. (2016). Enhancing literacy practices in science classrooms through a professional development program for Canadian minority-language teachers. *International Journal of Science Education, 38*(7), 1150–1173. https://doi.org/10.1080/09500693.2016.1183267
Romijn, B. R., Slot, P. L., & Leseman, P. P. M. (2021). Increasing teachers' intercultural competences in teacher preparation programs and through professional development: A review. *Teaching and Teacher Education, 98*. https://doi.org/10.1016/j.tate.2020.103236
Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal, 50*(5), 1020–1049. https://doi.org/10.3102/0002831213477680
Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. *Teaching and Teacher Education, 101*. https://doi.org/10.1016/j.tate.2021.103305
Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology, 6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016
Schneider, W., Baumert, J., Becker-Mrotzek, M., Kammermeyer, G., Rauschenbach, T., Roßbach, H.-G., Roth, H.-J., Rothweiler, M., & Stanat, P. (2012). *Expertise „Bildung*

durch Sprache und Schrift (BiSS)". Bund-Länder-Initiative zur Sprachförderung, Sprachdiagnostik und Leseförderung [Expertise "Education through language and literacy (BiSS)"]. n.p.. https://www.biss-sprachbildung.de/pdf/biss-website-biss-expertise.pdf

Seah, L. H., & Silver, R. E. (2018). Attending to science language demands in multilingual classrooms: A case study. *International Journal of Science Education, 42*(14), 2453–2471. https://doi.org/10.1080/09500693.2018.1504177

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189x015002004

Stanat, P., Becker, M., Baumert, J., Lüdtke, O., & Eckhardt, A. G. (2012). Improving second language skills of immigrant students: A field trial study evaluating the effects of a summer learning program. *Learning and Instruction, 22*, 159–170. https://doi.org/10.1016/j.learninstruc.2011.10.002

Steffensky, M., Gold, B., Holdynski, M., & Möller, K. (2015). Professional vision of classroom management and learning support in science classrooms – Does professional vision differ across general and content-specific classroom interactions? *International Journal of Science and Mathematics Education, 13*(2), 351–368. https://doi.org/10.1007/s10763-014-9607-0

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. *Journal of Research in Science Teaching, 39*(8), 664–687. https://doi.org/10.1002/tea.10040

Studhalter, U. T., Leuchter, M., Tettenborn, A., Elmer, A., Edelsbrunner, P. A., & Saalbach, H. (2021). Early science learning: The effects of teacher talk. *Learning and Instruction, 71*. https://doi.org/10.1016/j.learninstruc.2020.101371

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education, 3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1

Thomassen, W. E., & Munthe, E. (2020). A Norwegian perspective: Student teachers' orientations towards cultural and linguistic diversity in schools. *Teaching and Teacher Education, 96*. https://doi.org/10.1016/j.tate.2020.103151

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Ministry of Education.

Tong, F., Irby, B. J., Lara-Alecio, R., Guerrero, C., Tang, S., & Sutton-Jones, K. L. (2018). The impact of professional learning on in-service teachers' pedagogical delivery of literacy-infused science with middle school English learners: A randomised controlled trial study in the U.S. *Educational Studies, 45*(5), 533–553. https://doi.org/10.1080/03055698.2018.1509776

Vock, M., Gronostaj, A., Grosche, M., Ritterfeld, U., Ehl, B., Elstrodt-Wefing, N., Kalinowski, E., Möhring, J., Paul, M., Starke, A., & Zaruba, A. (2020). Das Projekt „Förderung der Bildungssprache Deutsch in der Primarstufe: Evaluation, Optimierung & Standardisierung von Tools im BiSS-Projekt" (BiSS-EOS) – Ergebnisse und Erfahrungen aus drei Projektjahren. [The project "Fostering academic language in German in primary school: Evaluation, optimization & standarization of tools in the BISS-project" (BISS-EOS) - results and experiences from three years of project]. In S. Gentrup, S. Henschel, K. Schotte, L. Beck, & P. Stanat (Eds.), *Sprach- und Schriftsprachförderung gestalten: Evaluation von Qualität und Wirksamkeit umgesetzter Konzepte*. Kohlhammer.

Wasik, B. A. (2010). What teachers can do to promote preschoolers' vocabulary development: Strategies from an effective language and literacy professional development coaching model. *The Reading Teacher, 63*(8), 621–633. https://doi.org/10.1598/rt.63.8.1

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yang, H. (2020, 2020/10/19). The effects of professional development experience on teacher self-efficacy: Analysis of an international dataset using bayesian multilevel models. *Professional Development in Education, 46*(5), 797–811. https://doi.org/10.1080/19415257.2019.1643393