International Journal of Teaching and Learning in Higher Education 2020, Volume 32, Number 2, 225-239
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Promoting and Measuring Reflective Skills in Depth and Breadth of English and Physics Teacher Trainees

Christiane Klempin and Daniel Rehfeldt
Freie Universität Berlin

Undoubtedly, supporting reflection in student teachers during university-based training is one of the most sustained measures to attain teacher professionalism. Therefore, at the Freie Universität Berlin an on-campus seminar designed to relate theory to practice and vice versa – the so-called *Teaching and Learning Lab (TLLS)* – was implemented over the course of five terms to stimulate reflective skills of English and physics teacher trainees. Investigations on the effectiveness of three types of the *TLLS* (no video and two types of video-supported reflections) compared to a parallel group (PG) and a control group (CG) occurred in a Mixed Methods quasi-experimental study. Reflective skills were elicited with vignettes, relevant covariates with questionnaires. Reflective development was then traced in the dimensions depth and breadth employing a Qualitative Content Analysis. MANCOVA and regression analyses revealed a substantive increase of reflective depth for English and physics teacher trainees and breadth development for English *TLLS*-participants in contrast to both, a PG and a CG, even when controlling for the subjects' individual prerequisites.

Introduction

Conceptualizing Reflection

Reflection is believed to contribute to "self-consciousness [which then] generates valid knowledge" (Fendler, 2003, p. 17). Dewey (1933, p. 9) regards it as "the active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it." Reflection herein is considered an active practice and a means to diminish an individual's unwarranted assumptions about teaching and learning. Any kind of passive action, in turn, is equated by Dewey with action based on "trial and error" (Griffiths, 2000, p. 540). Such routine and passive action is in danger of being justified by "authority and tradition" (Griffiths, 2000, p. 540). It is further hypothesized by Dewey that such rational reflection-rooted practice emerges in teachers when their observational and analytical skills have been trained systematically. Besides that, he considers refinement of three key personality features crucial for successful reflection: "open-mindedness, wholeheartedness and responsibility" (Griffiths, 2000, p. 540; cited in Dewey, 1933)

Schön (1983, p. 26) draws a distinction between "reflection-in-action" as it occurs intuitively in the moment of action and "reflection-on-action" as a retrospective act temporarily separated from the teaching act. This typology was supplemented with the term "reflection-for-action" coined by Körkkö, Kyrö-Ämmälä, and Turunen (2016, p. 200) to demonstrate that reflection may also take place before action, for instance, when planning for alternative action. Clarà (2015) rejects many of those widely accepted assumptions on reflection. For him, Schön's (1987) concept of *reflection-on-action* must not necessarily culminate in decision-making or alternative action plans. Rather, it should be considered a means to clarify pedagogical situations in order to gain greater insight and discernment (Clarà, 2015). Consequently, Clarà (2015, p. 263) defines reflection as "a thinking process which gives coherence to a situation which is initially incoherent and unclear." For this study, reflective performance will be expressed in accordance to Schädlich (2015, p. 258, translation and adaptation C.K.):

> Teacher trainees are capable to plan and conduct [English and physics activities] on the basis of subject-matter or curricular texts (theory). These teacher trainees are able to discuss the relevance of these texts for individual and complex experiences in a situation of action (praxis) and manage to explicate those field experiences. Reflective competence becomes explicit in trainees' performance to retrospectively verbalize their actions […] and their learning processes in the whole.

Promoting and Modeling Teacher Trainees' Reflection

Promoting Teacher Trainees' Reflection Skills

Claims are frequently made that the professionalization of teacher trainees largely rests on reflective teacher education (e.g., Davis, 2006; Leonhard, Wüst, & Helmstädter, 2011; Santagata & Yeh, 2013). Literature review reveals five recurring objectives associated with a teacher education promoting reflective practices. Reflection is regarded to stimulate:

I. The process of becoming conscious of one's own beliefs on teaching and learning which transforms implicit (unconscious, intuitive) to explicit (rational, theory-guided) knowledge

(e.g., Körkkö, Ämmälä, & Turunen, 2016; Häcker, 2017).
II. The negotiation of the relation between knowledge and performance (e.g., Bullough, 1989; Day, 1993).
III. The negotiation of the theory-practice-relation (e.g., Wildt, 1995; 2003; Birmingham, 2004).
IV. Dealing with pedagogical dilemmas (e.g., Combe & Kolbe, 2008; Leonhard, Wüst, & Helmstädter, 2011).
V. Rethinking and innovating the educational landscape (e.g., Schön, 1987; Smyth, 1989; Bullough, 1989; Day, 1993; Dewey, 1933).

Developing reflective skills serves as one of many means to attain the features of a professional educator. In order to bring this important skill closer to student teachers the *Teaching and Learning Laboratory (TLLS)* was constructed implementing specific micro-interventions to promote reflective skills in a training environment combining theory and practice elements. According to Abendroth-Timmer (2017), four approaches can be differentiated in teacher education to promote reflective reasoning, all of which were implemented in the underlying intervention, the *TLLS*:

I. Individual-monological approaches include self-reports, such as journals (Rieger, Radcliffe, & Doepker, 2013), diaries (Akbari, 2007), or portfolios (Schädlich, 2015; Abendroth-Timmer, 2017).
II. Collegial-dialogical approaches rest on peers as "critical friend[s]" (Hatton & Smith, 1995, p. 37) or experts as "knowledgeable others" (Gelfuso & Dennis, 2014, p. 2).
III. Visual approaches allow for critical reflection-on-action (Schön, 1983) of teacher trainees' own teaching (personal video reflection), or other's teaching efforts (external video reflection). Both of these video reflections – personal and external – were realized in the *TLLS*.
IV. Experimental approaches comprise action research projects (Legutke & Schart, 2016), simulations (Abendroth-Timmer, 2017), blended-learning experiences (Abendroth-Timmer & Frevel, 2013), as well as Microteaching (Schädlich, 2015)

Modeling Teacher Trainees' Reflective Skills

In the past, products of student reflections were empirically evidenced to vary with respect to their quality. Assessment of "reflective depth" as one alleged sub-dimension of reflective skills (Leonhard, Wüst, & Helmstädter, 2011), for instance, occurred with reference to hierarchical reflectivity models (Hatton & Smith, 1995; Abels, 2011). These models were employed to reconstruct various reflectivity levels from the reflective data material (verbal, written, or videotaped).

Hatton and Smith (1995) laid foundations for the four-level-model of reflection[1] utilized by likewise, the authors of this contribution and Abels (2011). Abels (2011) adapted Hatton's and Smith's model to determine teacher trainees' reflective performance in written portfolios. The research project discussed in this paper relies on the four-level-reflection-model by Abels (2011).

Investigation of "reflective breadth" as a second dimension of "reflective skills" can yield valuable additional information on the quality of teacher reflections (Leonhard et al., 2011) and can be operationalized by *Pedagogical Content Knowledge (PCK).* According to Shulman (1987, p. 9), teachers' *PCK* translates *Content Knowledge (henceforth CK)* from English Linguistics, Literary, and Cultural Studies into "comprehensible" input for language learners. This specific characteristic ultimately renders *PCK* a "special amalgam of content and pedagogy" (Shulman, 1987, p. 9).

Very limited research was conducted on *PCK* in English Language Teaching (ELT; e.g. Akbari & Tajik, 2009; Freeman & Johnson, 1998; Gatbonton, 1999). Systematic, large-scale research projects have only been realized very recently. As one example, in the *PKE*-study (*Professional Competencies for English Teacher Trainees* (Kirchhoff, 2017), Shulman's three knowledge domains were used as a framework by König and colleagues (2016, p. 322) to track trainees' *PCK*, *PK*, *General Pedagogical Knowledge (GPK)* and *CK* development over the course of their university-based training. The *PCK* knowledge assessment tool employed in *PKE* comprises the subsequent three dimensions (Darling-Hammond & Baratz-Snowden, 2007; Jansing Haudeck, Keßler, Nold, & Stancel-Piatak, 2013):

1. Curriculum knowledge (CURR)
2. Strategy and representational knowledge (STRAT)
3. Learner Knowledge (LEARN)

König and colleagues (2016) statistically confirmed a *PCK*-model which subdivides CURR- and STRAT-knowledge into six and LEARN-knowledge into another seven domains. This *PCK*-trias was also deployed in this study.[2]

[1] By way of analyzing the content of teacher trainees' reflection essays Hatton and Smith (1995, p. 41) identified four types of reflection: *I. descriptive writing, II. descriptive reflection, III. dialogic reflection*, and *IV. critical reflection*.

[2] This project is part of the "Qualitätsoffensive Lehrerbildung", a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The programme is funded by the Federal Ministry of Education and Research. The authors are responsible for the content of this publication

In sum, the *PKE*-study evidenced that relationships between the domains resulted in either high (*CK* vs. *PCK*; *GPK* vs. *PCK*) or medium correlations (*GPK* vs. *CK*). As compared to a stronger correlation between *CK* and *PCK* for mathematics teaching in the *COACTIV* study (r = .78), the role of *CK* in English Language Teaching (ELT) appears to differ. Taking into consideration the role of English in *ELT*, being a simultaneously learning vehicle and an objective of *Teaching English as a Foreign Language (TEFL)*, this finding is hardly surprising.

Treatment Description: *The TLLS*

The *TLLS*-format was designed and implemented as a quasi-experimental teacher training approach" in the subject-matter domains of English and physics according to a standardized theoretical framework (Rehfeldt, Seibert, Klempin, Mehrtens, Lücke, Sambanis, Köster, & Nordmeier, 2018). For the purpose of developing teacher trainees' reflective skills, implementation of three micro-interventions derived from the body of literature on effective strategies to foster reflection occurred. The *TLLS* blends elements of a regular theory seminar with iterative field explorations with pupils, followed by reflective sessions.

In a *TLLS*, teacher trainees develop theory-grounded learning opportunities which are first put into practice with visiting learners in university spaces. Field experiences are then subject to reflections, theory-based optimization, and iterative explorations with learners.

The conceptual framework of the *TLLS* traces back to the process model developed by Nordmeier and colleagues (2014). The *TLLS*-conception model describes five main steps, whereby steps (b) and (d) are repeated:

(a) Pre-selected theory input and supported planning of instruction (approx. 6 sessions)
(b) Conduction and exploration of instruction plan (1 session)
(c) Theory-based observation of peer exploration (occurs during b)
(d) Theory-based reflection of field experience (1 session)
(e) Theory-based adaption and modification of instruction plan (approx. 1-2 sessions)

In order to actively guide teacher students and support their skill development, a competence model targeting expert teacher perception and action skills, was employed and is further discussed in Klempin (2019).

Micro-Interventions to Foster Reflective Skills

At the Department of English subject-matter education at *Freie Universität Berlin*, the *TLLS* was specifically implemented at the B.A. level for the promotion of English teacher trainees' didactic reflective skills (Klempin, 2019). At the Department of Physics Teacher Education, the same concept model was employed, including the three micro-interventions to foster reflective skills of the participants. This has enabled the researchers to conduct joint data analyses.

First, a Cognitive Apprenticeship (Schädlich, 2015) based on the instructor's model was pursued, specifically during those phases which relied on theory input by the instructor (a, d & e). As a second micro-intervention, *Noticing Trainings* (Sherin & van Es, 2009) were carried out. For that, trainees were to collaborate with "critical friend[s]" as advised by Hatton and Smith (1995, p. 37) during instruction planning (phase a), field explorations (phase b), and peer observation (phase c). For further advancement of students' observation skills, pre-structured protocols were to be filled out during field explorations (phase b). Along with short teaching video clips, these observation notes were used to stimulate participants' theory-guided reflections (phase d).

The two highly structured reflective sessions, framing both praxis phases of the *TLLS*, were modeled on the reflective cycle by Rodgers (2002). This reflection-on-action (Schön, 1983) provides trainees with opportunities to look back onto their praxis experiences while making references to relevant theories when passing through the steps of the cycle. Reflection occurs individually as well as in peer tandems and with the mentoring instructor providing assistance whenever required (Kaasila & Lauriala, 2012). Following reflection, the activities were revised for performance improvement throughout the second field trial. These revisions are based on students' findings from the first reflective session, peer feedback, and observation protocol notes the tandem partner took during the activity. It is assumed that through the reflective sessions participants acquire the epistemic skills to reminisce on their field experiences in a theoretically sound fashion. This could, for instance, be realized by way of imagining alternative paths of action for the second field trial (phase e).

Research Hypotheses

Our research project is driven by the assumption that didactic skills of participants of the *TLLS* will extend with regard to their reflective depth and breadth. Reflective depth is operationalized as modes of perspectives and analyzed employing a reflectivity model (Abels, 2011). Reflective breadth is hereby approached as the display of *PCK* (König, Lammerding, Nold, Rohde, Strauß, & Tachtsoglu, 2016). Further, we hypothesize different effects on reflective depth and breadth development depending on the type of intervention participants were exposed to.

Three main *TLLS*-interventions are distinguished, all comprising the aforementioned micro-interventions to

support reflection (3.1). However, interventions II and III also include the visual approach integrating either trainees' own short teaching video clips ("personal video"), or other candidates' video clips ("external video") into the reflective sessions. Drawing from prior research findings (Santagata, Zannoni, & Stigler 2007, p. 344), we expect the highest increase of reflective depth in intervention group II, followed by III and I:

I. Intervention "regular *TLLS*": *TLLS* with 3 micro-interventions
II. Intervention "personal video *TLLS*": Regular *TLLS* plus videography and reflective sessions with personal videography
III. Intervention in "external video *TLLS*": Regular *TLLS* without videography but with reflective sessions with someone else's videography

This research project pursues the following research hypotheses on reflective skill development of *TLLS*-participants compared to subjects from a parallel (PG) and those from a control group (CG).

Research Hypotheses on Reflective Depth Development. H1: The reflective depth development differs measurably between participants of the five intervention types (regular *TLLS*, parallel *TLLS*/PG, control group/CG and *TLLS* with either personal or external video reflection) with a medium effect size in favor of the *TLLS*-participants and even with a high effect size resulting from the additional video-supported reflection (Bandura, 1997, p. 79-81; Tschannen-Moran, Hoy, & Hoy, 1998; Schmitz, 2000, p. 16; Helbig, Günther, Rehfeldt, & Krüger, 2019).

Research Hypotheses on Reflective Breadth Development. H2: The reflective breadth development differs measurably between participants of the five intervention types (regular *TLLS*, parallel *TLLS*/PG, control group/CG and *TLLS* with either personal or external video reflection) with a medium effect size in favor of the *TLLS*-participants and even with high effect size resulting from the additional video-supported reflection.

Further Research Hypotheses. H3: According to a theoretical framework, the statistical relation between the increase of reflective depth and breadth is positive and medium to strong (representing an overall reflection skill/competence).

Research Design, Methods, and Instruments

The *TLLS* in the subject-matter education of English and physics are explored within the research paradigm of Mixed Methods (Kuckartz, 2014) as to pay tribute to the complexity of the construct under investigation (Häcker, 2017). "Didactic reflection", in its alleged dimensions "reflective depth" (Hatton & Smith, 1995; Abels, 2011; Leonhard, Wüst, & Helmstädter 2011; Roters, 2013; Stender, 2015) and "reflective breadth" (König, Lammerding, Nold, Rohde, Strauß, & Tachtsoglu, 2016), was elicited at pre- and post-points of measurement. Data was collected in five experimental cohorts (interventions I-III), one parallel *TLLS* (PG follows *TLLS*-concept but lacks the three micro-interventions), and three content-wise comparable control groups (CGs constitute regular theory courses without field practice and systematic reflective sessions). Pre-test data collection always occurred in the initial week of the term, whereas post-test data was collected during the penultimate or ultimate session. Table 1 gives insight into the distribution of additional interventions in the *TLLS*.

Data on reflective depth and breadth was elicited using open written discourse vignettes which were developed based on a theoretical framework (Rehm & Bölsterli, 2014[3]). Additionally, a paper-pencil questionnaire was distributed in EG, PG, and CG at pre and POST points of measurement to allow for covariate analyses of such variables theoretically suspected to affect the development of reflective skills.

Analysis

Study I: Reflective Depth and Breadth Development of *ELT* Trainees

Sample Description Study I

This intervention study is based on a quasi-experimental design whereby participation in the intervention seminars was voluntary. All five interventions were realized at the Freie Universität Berlin (Germany). The *TLLS* were conducted as part of teacher training during the Bachelor's program spanning summer terms 2016 to 2018. The cohorts which were used for further analyses comprised English and physics teacher trainees. Samples for the *TLLS* without video support were recruited from the English cohorts of summer term 2016 to winter term 2016/17. Control groups could be established during the summer terms from 2016 until 2018. A parallel *TLLS* (PG) exists for English only and was conducted by a fellow researcher during summer term 2018. *TLLS* with personal video occurred during summer term 2017 in English and the external video reflection was implemented in the *TLLS* offered during winter term 2017/18. More detailed information on the samples and sub-samples is provided in table 2.

[3] For more insight into the construction of the *open written discourse vignettes*, see Klempin (2019).

Table 1
Distribution of Experimental Groups (EG/TLLS), Parallel Group (PG), and Control Groups (CG) From Summer Term 2016 Until Summer Term 2018 in English and Physics

Subject	Summer 2016	Winter 2016	Summer 2017	Winter 2017	Summer 2018
English	Intervention I: regular *TLLS*	Intervention I: Regular *TLLS*	Intervention II: Personal Video	Intervention III: External Video	Intervention I: Regular *TLLS*
	Control group (CG)		Control Group (CG)		Parallel Group (PG)
					Control Group (CG)
Physics					Intervention II: Personal Video

Table 2
Comparison of the Samples of the Five Interventions (Regular TLLS, Personal TLLS, External TLLS, Parallel Group (PG), and Control Group (CG)) According to Age, Semester, Gender, Practical Experience, and A Level Sum Score (Sum Score of All Grades Granted for the German A Levels. A Levels Are a Type of General Qualification for University Entrance.). The Practice-Score Ranges from 0: No Prior Experience up to 5: Rich Prior Practical Experience as Student Teacher or Private Tutor etc.

Intervention	Age (SD)	Semester (SD)	Gender	Practical Experience (SD)	A Level Sum Score (SD)
CG	24.1 (3.82)	4.1 (1.26)	16f, 5m	1.33 (1.46)	2.1 (0.64)
PG	21.42 (1.56)	4.58 (1.24)	8f, 4m	1 (0.6)	2.43 (0.78)
TLLS	22.17 (2.98)	4.07 (1.16)	23f, 6m	0.72 (0.88)	1.94 (0.54)
TLLS external	23.06 (3.86)	3.94 (1.43)	8f, 10m	2.56 (1.42)	2.13 (0.44)
TLLS personal	21.83 (2.32)	3.67 (1.37)	4f, 2m	3 (1.1)	2.07 (0.84)

With regard to those variables used to describe the sample (see Table 2), differences were found to be statistically insignificant when employing a MANCOVA (dependent variable/DV: reflective breadth and reflective depth, independent variable/IV: time, factor: intervention type). For subsequent analyses, therefore, samples of the five groups are treated as comparable.

Qualitative Analysis for Reflective Depth and Breadth

Qualitative Analysis for Reflective Depth. For analysis of English teacher trainees' "reflective depth", an inductive Qualitative Content Analysis (Kuckartz, 2016) was pursued to gradually extract a four-category-system from the data output. The assessment tool MAXQDA Plus 12 was employed for data coding. Categories were coined inductively in a double-blind coding process by the *TLLS*-instructor-researcher, as well as by two trained student assistants (one in each subject). During the initial rounds of analyses (cohort of summer term 2016), disagreements between the coders were discussed until mutual agreement was reached (Kuckartz, 2016). Further, the category system derived during this very first phase was inspected and consensually negotiated by a group of subject-matter education experts from English, physics, primary education, and history. After coding, the vignettes were subject to double-blind allocation by two raters (for each subject) to four levels of the adapted reflectivity model by Abels (2011). Formative modifications on the original model by Abels occurred as a result of the inductive coding process of the first phase which led to the realization that alterations were required due to diverging subjects, data material, and finally content. Model adaptations gradually emerged in all phases of the analytical processes (summer term 2016 to summer term 2018) until eventually the subsequent modes of reflective depth appeared from the data:

1. Descriptive mode
2. Hypothesis mode
3. Explorative-productive mode
4. Multi-perspective-productive mode

Reflective modes (1-4) assigned to all teacher trainees' vignettes were then transformed into rank scores, giving way for inferential statistics. According to Krippendorff (2004), intercoder reliability indicated a good overlap across all times of measurement, spanning all cohorts from summer term 2016 to summer term 2018 ($\alpha = .92$***).

Qualitative Analysis for Reflective Breadth. The quality of a reflection on teaching also largely depends on how much knowledge someone has at his/her disposal. If reflection now occurs on a subject-specific teaching issue – for instance, on how to handle students' misconceptions on climate change in physics or how to initiate and support communication in English – the quality of the reflection becomes overt in a person's ability to retrieve *PCK* in such a way as to clarify the teaching problem (Roters, 2012, p. 387).

Since *PCK* has been proven to be highly topic-specific, the open written discourse vignettes utilized to elicit reflection in this study proved helpful in visualizing the knowledge teacher trainees referred to when trying to make sense of the very subject-specific teaching problem which was presented in the impulse. This reverberates the notion of reflection as an act of establishing situational coherence (Clarà, 2015) and as a very specific type of problem-solving (Berliner, 2004). The more and varied the display of *PCK* in the reflective output, the higher our estimation of the quality of the reflection.

Analyzing English teacher trainees' "reflective breadth", a deductive Qualitative Content Analysis (Kuckartz, 2016) was utilized. Categories were coined in a deductive fashion referring to the three *PCK*-dimensions deployed in the *PKE*-study by König and colleagues (2016). The deductive content analysis was conducted by coders working double-blind and ignorant to both the intervention group and the time of measurement. Quality of the coding processes was ensured through several measures following recommendations by Steinke (2007) and Kuckartz (2016). After each coding loop, the double-blind codings of the two researchers were discussed in a team of three coders until mutual agreement on the adaptations and refinements made on the category system was reached (Kuckartz, 2016). The final category system was critically inspected and approved by a team of *ELT* experts.

The formative coding process resulted in slight adaptations of the original *PCK*-model by König and colleagues (2016). In sum, two major changes in the category system with regard to the first *PCK*-domain CURR-Knowledge occurred. First, the sub-dimension *TEFL* was too unspecific to allow for distinct coding. Therefore TEFL was theoretically further specified by consulting Legutke and Schart (2016, 18-20). This yielded the novel *PCK*-domain TEFL compartmentalizing into the four sub-domains: 1. Linguistic, literary, and cultural knowledge; 2. Knowledge about teaching and learning; 3. Identity and role development; and 4. Cooperation and professional development.

Secondly, the sub-domains "learning goals" and "development goal" (both CURR) were not only impossible to distinguish in the coding process, but were also found to lack theoretical foundation in the respective *ELT* literature (i.e., German educational plans and framework curricula). Consequently, these domains were merged into one sub-domain called "development and learning goals". The following final category system emerged from the deductive coding process:

1. TEFL knowledge (TEFL)
2. CURRICULUM knowledge (CURR)
3. STRATEGY knowledge (STRAT)
4. LEARNER knowledge (LEARN)

Quantitative Analyses for Reflective Depth and Breadth of *ELT* Trainees

In this first study, missing values were treated with multiple imputation, proceeded by pool-procedure (Van Buuren & Groothuis-Oudshoorn, 2011). If imputation was impossible (5%-criterion, Van Buuren & Groothuis-Oudshoorn, 2011), a listwise case exclusion was chosen. The two constructs "reflective depth" ("reflex") and "reflective breadth" ("reflbr") were elicited at two points of measurement, pre-test (before the intervention) and post-test (after the intervention). Under consideration of covariates, a multivariate analysis – specifically a MANCOVA with repeated measures – was selected due to the fact that the investigation of the development of two constructs as two dependent variables was pursued. We are aware that a structural equation model might have yielded greater accuracy. However, such procedures are inappropriate for our small samples ($N \ll 300$, Brown, 2006, p. 305). For MANCOVA, both constructs "reflex" and "reflbr" were defined as dependent variables, and the point of measurement (pre-test vs. post-test), as well as the intervention type (*TLLS*, *TLLS* personal, *TLLS* external, PG, and CG), were determined as factors. As covariates, trainees' prior practical experience was implemented to gauge its impact on reflective development, and the A level sum score was used to determine the cognitive prerequisites. Further, the personality traits for successful reflection according to Dewey (1933), the covariates prior reflection knowledge, gender, age, and semester, as well and character count of the reflection output at pre-point of measurement[4] were investigated. With a sample size of $N = 86$ for the

[4] Character count of the reflection output served to control for the effect of the method on the study participants' motivation to reflect in writing.

Table 3

Comparison of the Samples of the Five Interventions (Regular TLLS, Personal TLLS, External TLLS, Parallel Group (PG), and Control Group (CG)) According to Age, Semester, Gender, Practical Experience, and A Level Sum Score. The Practice-Score Ranges from 0: No Prior Experience up to 5: Rich Prior Practical Experience as Teacher Student or Private Tutor etc.

Intervention	Subject Distribution	Age (SD)	Semester (SD)	Gender	Practical Experience (SD)	A Level Grade (SD)
CG	27 English	23.80 (3.57)	4.30 (1.27)	21f, 6m	1.15 (1.38)	2.19 (0.63)
PG	14 English	21.93 (2.76)	4.50 (1.16)	9f, 5m	1.00 (0.68)	2.32 (0.85)
TLLS Regular	38 English	23.38 (4.92)	4.16 (1.21)	25f, 13m	0.63 (0.82)	2.00 (0.58)
TLLS External	18 English	23.94 (3.86)	4.00 (1.50)	8f, 10m	2.72 (1.32)	2.13 (0.44)
TLLS Personal	7 English, 8 Physics	25.87 (5.25)	6.07 (2.52)	8f, 7m	1.81 (1.37)	2.25 (0.71)

analysis of reflective breadth and depth in English via MANCOVA, a power of .80 and a level of significance of $\alpha = .05$, medium effect sizes can be resolved (f(V) = .27; Bortz, 2010, p. 481).

Results

Results of Study I: Reflective Depth of English and Physics Teacher Trainees

Data for study I overlaps notably with that of study II. However, in study II, physics teacher trainees were included in the dataset, along with those English teacher trainees who were omitted from the analyses of study I due to their missing data.

Sample Description Study II

The sample analysis is comparable to that of study I. The sample of study II is detailed in Table 3.

Qualitative Analysis of Reflective Depth for English and Physics. In terms of elicitation and investigation, reflective depth was analogously treated to study I. Methodically, the four-category coding scheme which was developed for English was transferred to these vignettes in order to determine the level of reflective depth achieved by the physics teacher trainees. Intercoder reliability for English and physics subject-matter education yields a good overlap across all times of measurement ($\alpha = .91$***).

Quantitative Analysis of Reflective Depth for English and Physics. In the second study missing values were again treated with multiple imputation, pool-procedure (Van Buuren & Groothuis-Oudshoorn, 2011), or listwise exclusion. Based on the identical conception of the *TLLS* in English and physics (Klempin, Rehfeldt, Seibert, Mehrtens, Nordmeier, Lücke, Köster, & Sambanis, 2019), despite the two distinct subject-matter domains, data for both was aggregated. An interdisciplinary analysis was realized as part of the covariate analyses (see for further information below), whereas covariates were equal to those in study I. Multiple linear regression analyses were employed to quantitatively investigate the development of "reflective depth (reflex)" of English and physics teacher trainees. For all regression analyses, the relevant statistical assumptions were tested and found to be met (normal distribution, homoscedasticity, normally distributed residuals, no multicollinearity). Exclusion of outliers was conducted for each model (Mahalanobis Distance, Cook's Distance, and Leverage with equal weighing; exclusion when a minimum of two out of three criteria was met) with an exclusion rate of < 6%.

For elaboration of H1, reflective depth at the end of the intervention (post-test) was defined as the dependent variable, whereas reflective depth at the beginning of the seminar (pre-test), as well as the intervention type (1: CG, 2: PG, 3: *TLLS* regular, 4: *TLLS* external, 5: *TLLS* personal), were determined to be the independent variables. This procedure enables to control for the pre-baseline, which in turn guarantees improved validity of the outcome. As another benefit, this method allows for implementation of interval-scaled covariates in order to gauge an estimate of means adjusted for covariates. The latter can be illustrated in the unstandardized regression coefficient of the independent variable "intervention type". Estimation of the adjusted mean of reflective depth can also be achieved by centering the means of the independent variable and covariates except for the control group. Mathematically, the determination of an adjusted effect size for the mean difference of reflective depth between the three *TLLS*-formats and the control and parallel group can be realized.

For H2, the dependent variable was defined as the difference between the reflective depth values of a subject at post-test and pre-test measurement point. As an independent variable, again, the "intervention type" was selected, albeit, now centered. This prior analysis serves to statistically ensure mean variation between entry (pre-test) and exit reflective depth (post). It also detects developmental differences regarding the intervention group. With a significant intercept (average mean differences "intervention type" vs. "reflective depth") it is statistically appropriate to test post-hoc for mean differences between the three *TLLS* interventions, the PG, and the CG without α-correction (Field, Miles, & Field, 2012, p. 745). A significant regression coefficient of the CG (centered) indicates differences in the perception of EG and CG.

H3 is investigated as part of the above-mentioned analyses via implementation of planned contrasts. All of the aforementioned analyses were calculated using Rstudio (lm()-package). With a sample size of $N = 112$ via regression analysis, a power of .80 and a level of significance of $\alpha = .05$ will resolve small to medium effects ($f^2 = .07$, Bortz, 2010, p. 481; Faul, Erdfelder, Buchner, & Lang, 2009).

Results

Study I Results: Reflective Depth and Breadth in English

Statistical assumptions were first tested for the MANCOVA (multivariate normal distribution, homogenous variance-covariance matrices; Bortz, 2010, p. 481). Both assumptions were found to be violated (Box-Test: $\chi^2(27) = 48.3$, $p = .007$; Shapiro-Wilk-Test: $W = 0.96$, $p < .001$), but MANCOVA is assumed to be a robust procedure for sample sizes exceeding 30 (Allen & Bennett, 2007). The A level sum score, gender, "prior knowledge on the reflection of teaching," "prior teaching experiences," and the character count (pre) were implemented as covariates. In sum, all of these covariates were statistically insignificant ($ps > 0.05$). Besides that, the impact of the "personality traits for successful reflection" yielded insignificant values ($ps > 0.05$). However, significant effects were detected for both factors, the time of measurement, and also the intervention type. In particular, the significance of the effect of the interaction term time*intervention (see Table 5) is worth mentioning, as this might serve as a first indicator for different developments of reflective depth and breadth in the five intervention types. It was then proceeded with univariate ANOVAs (Field, Miles, & Field, 2012, p. 745). The subsequent univariate ANOVAs consolidate and diversify the insights gained from the earlier MANCOVA analysis (Tab. 3). Time of measurement, intervention type, and the interaction term time*intervention type remain significant. This suggests differences for both dependent variables across time and with respect to the intervention format.

Post-hoc tests were conducted afterwards. Therefore, the progression of either constructs was modelled as a pre- and post-differential measure, whereby positive values indicate a positive pre- and post-development. If one tests these differential measures, contrasting the interventions via t-test or Mann-Whitney-U-Test[5], significant differences between the interventions *TLLS*, *TLLS* personal video, as well as *TLLS* external video against the PG and the CG are yielded ($ps < .001$). Contrasting all *TLLS*-interventions with the CG and the PG significant and medium effects for both dependent variables, in favor of the *TLLS*-interventions ($ds > 0.67$), are evidenced. Between the three *TLLS*-interventions no significant differences could be detected ($ps_{depth} > .051$, $ps_{breadth} > .56$, $p = .47$):

For the increase in reflective depth an investigation with respect to the sub-dimensions was pursued (see Table 6). It is evidenced that significant differences solely occur in some comparisons of the CG and the PG with the *TLLS*-formats, whereas the *TLLS*-interventions do not differ with regard to the reflective breadth development per sub-dimension. For the sub-dimension STRAT-Knowledge in all *TLLS*-formats differences are statistically significant compared to PG and CG. The TEFL-dimension develops significantly more exclusively in the regular *TLLS*, again compared to PG and CG. In contrast, the LEARN-dimension yields significant increases only for the *TLLS* with external video reflections. The effect sizes range from medium to high. For the CURR-dimension no differences were detected.

When considering global statistical efficiency (overall reflective depth and breadth), the five interventions, which were originally assumed in this paper, result in solely two intervention types. As a result, CG and PG can be considered one intervention type whereas all *TLLS*-types (regular, external, and personal) can be counted as a second treatment group. Merely on the level of the sub-dimensions of reflective breadth, specific differences were identified. According to the final H3, a correlation of $r = .51$ ($p < .001$) between the increase of reflective depth and breadth was detected.

[5] The moderate to small sample sizes per intervention were approached with a parameter-free test.

Table 4

Results of MANCOVA. Dependent Variables are "Reflective Depth" and "Reflective Breadth". Factors are the Point of Measurement (Time) and the Intervention Type (Intervention).

Factor/Covariate		Test Statistic	F	*df1*	*df2*	*p*
Time	Pillai's Trace	0.185	18.52	2	163	< .001
	Hotelling's Trace	0.185	18.52	2	163	< .001
Intervention	Pillai's Trace	4.87	4.87	*8*	328	< .001
	Hotelling's Trace	5.28	5.28	*8*	324	< .001
Time*Intervention	Pillai's Trace	3.68	3.68	*8*	328	< .001
	Hotelling's Trace	3.66	3.66	*8*	324	< .001

Table 5

*Results of the ANOVAs. Dependent Variables are "Reflective Depth" (reflex) and "Reflective Breadth" (reflbr). Factors Are the Time of Measurement (Time) and the Intervention Type (Intervention). A Significant Effect was Evidenced for the Measurement Time (Time), the Intervention Type as well as the Interaction Between Measurement Time and Intervention (Time*Intervention).*

	Dependent Variable	Sum of Squares	*df*	Mean Square	*F*	*p*
Time	reflex	14.37	1	14.37	14.37	< 0.001
	reflbr	0.38	1	0.38	0.38	< 0.001
Intervention	reflex	11.34	4	2.84	2.84	< 0.001
	reflbr	0.94	4	0.23	0.23	< 0.001
Time*Intervention	reflex	6.92	4	1.73	1.73	< 0.001
	reflbr	0.45	4	0.11	0.11	< 0.001
Residuals	reflex	64.33	164	0.39	0.39	< 0.002
	reflbr	4.92	164	0.03	0.03	< 0.006

Table 6

Analysis of Significant Mean Differences for Reflective Breadth and Reflective Depth Depending on the Intervention Type.

Reflective Breadth (Increase)					
Intervention 1	Intervention 2	M1	M2	*p*	*d*[6]
CG	*TLLS*	0.31	0.53	< 0.001	0.87
CG	*TLLS*+External	0.31	0.58	< 0.001	0.95
CG	*TLLS*+Personal	0.31	0.56	0.006	0.85
PG	*TLLS*	0.28	0.53	< 0.001	0.98
PG	*TLLS*+External	0.28	0.58	< 0.001	1.11
PG	*TLLS*+Personal	0.28	0.56	0.003	1.08

Reflective Depth (Increase)					
Intervention 1	Intervention 2	M1	M2	*p*	*d*[7]
CG	*TLLS*	0.14	0.86	0.002	0.73
CG	*TLLS*+Personal	0.14	1.50	0.000	1.04
PG	*TLLS*	0.08	0.86	0.005	0.74
PG	*TLLS*+Personal	0.08	1.50	0.001	1.17
+External Video	*TLLS*+Personal	0.61	1.50	0.05	0.67

[6] Only for the comparison of CG vs. *TLLS* a classical independent t-test was calculated, whereas all other comparisons were realized via U-test. The classical effect size *r* was calculated referring to Field (2012, p. 665) and converted into Cohens *d*.

[7] Only for the comparison of CG vs. *TLLS* a classical independent t-test was calculated, whereas all other comparisons were realized via U-test. The classical effect size *r* was calculated referring to Field (2012, p. 665) and converted into Cohens *d*.

Results study II: Reflective Depth in English and Physics for Each Intervention

To begin with, differences in the increase of reflective depth per subject (English vs. physics) were insignificant (t-test, $p > .73$), so data was aggregated. The linear regression (see Table 7) of the teacher trainees' reflective depth at the end of the seminar (post) as a dependent variable, along with the reflective depth at the beginning of the course (pre) and the intervention type as an independent variable, resulted in a significant prediction model ($F(5, 106) = 13.76$, $p < .001$, $R^2 = 0.36$) with a significant intercept ($I = 2.38$, $SE = 0.11$, $p < .001$), a significant regression coefficient for the reflective depth of students at the start of the seminar ($b = 2.38$, $SE = 0.11$, $p < 0.002$, $\beta = .26$), as well as partially significant contrasts of the intervention types against the CG (see Table 7).

It is noteworthy that only the contrast CG vs. PG turns out to be insignificant ($p = .76$), and all remaining contrasts are highly significant to the disadvantage of the CG ($\beta s > 0.34$). The covariates were not further incorporated into the analysis, as the correlations with the target variable were found to be small and, hence, negligible ($rs < .22$). The independent t-tests and Mann-Whitney U-tests ($N < 20$, for at least one group), which investigate the different intervention types, could not evidence any differences in means for reflective depth increase across the *TLLS*-formats ($ps > .26$). Additionally, there are no differences between both the PG and the CG ($p = .76$). Table 8 shows the significant results, including the effect sizes. However, significant differences with medium to high effects could be ascertained for all *TLLS* formats as compared to both the PG and the CG, especially the differences in means for the increase in reflective depth between the *TLLS*-formats and the CGs, as well as the PG yield statistical significance ($ps < .02$), mostly even with high effects.

Discussion

Discussion Study I: Reflective Breadth Development

Concerning H2 we are able to report significant differences for some comparisons of the PG as well as the CG against the *TLLS* formats. Simultaneously, the *TLLS* formats do not differ statistically in terms of reflective breadth, even per sub-dimension. Hence, the first part of the initial research hypothesis 2 can be confirmed. For the second part, the data is somewhat inconclusive and does not suffice to claim that reflective breadth development increases significantly more with video-supported reflection in the *TLLS*. So far, it appears that the regular *TLLS* with its three explicit micro-interventions to support reflection avails the promotion of an overall reflective breadth development in teacher trainees, regardless of additional video reflections.

Taking a look at the particular dimensions of *PCK*, STRAT-Knowledge displays a significantly higher increase across all *TLLS*-formats contrasted to both, the PG and the CG. Conversely, TEFL-knowledge only yields a measurable enhanced increase for the regular *TLLS* as contrasted to both the PG and CG. On the other hand, LEARN-knowledge develops much stronger with medium to strong effect sizes exclusively for the *TLLS* with external video reflection compared to CG and PG. No significant differences between reflective depth developments for the five interventions can be reported for CURR-knowledge.

Based on these findings one might assume that watching someone else's teaching performance – as part of the *TLLS* intervention with external video reflection – rather focuses teacher trainees' attention on the learner. One should, however, take into consideration that LEARN-knowledge has already been addressed quite frequently at the beginning of the course. Consequently, this shift in attention might be an indicator that the focus on the learner remains stable or is even enhanced only when reflection is supported by external videos.

CURR-knowledge did not increase regardless of the *TLLS*-type, which might be explained by the location of the module (Bachelor's program). Curriculum does not yet play such a crucial role there.

In this study a pronounced development in the sub-domain LEARN-knowledge was found for participants of the *TLLS* as compared to teacher trainees who attended the non-*TLLS*-formats. Experienced teachers were evidenced to target the learning and apprehension processes of learners in their reflections (Borko & Livingston, 1989). Those teacher trainees who participated in the external video intervention display such skills, as they sustain and extend their learner focus over the span of the course while also attempting to develop appropriate strategies to support student learning. According to Neuweg (2007, p. 94), expertise is expressed by teachers' unconditional orientation towards the learner. Such behavior is assumed to increase the probability for a context-sensitive perception and diagnosis and in turn, meaningful and student-centered teacher actions (Neuweg, 2007, p. 94). We see such an orientation reflected in *TLLS*-participants' overall increased explication of STRAT-knowledge and LEARN-knowledge, the latter being valid solely for *TLLS* external video though.

Table 7
Breakdown of Significant Reflective Breadth Increases According to Dimension.

Intervention 1	Intervention 2	M1	M2	*p*	*d*[8]
Reflective Breadth (Increase): TEFL					
CG	*TLLS*	0.18	0.44	0.002	0.74
PG	*TLLS*	0.23	0.44	0.03	0.57
Reflective Breadth (Increase): LEARN					
CG	*TLLS*+External Video	0.36	0.67	0.02	0.64
PG	*TLLS*+External Video	0.32	0.67	0.02	0.72
Reflective Breadth (Increase): STRAT					
CG	*TLLS*	0.48	0.78	0.03	0.53
CG	*TLLS*+Personal Video	0.48	1.03	0.006	0.85
PG	*TLLS*	0.35	0.78	0.01	0.67
PG	*TLLS*+External Video	0.35	0.82	0.05	0.62
PG	*TLLS*+Personal Video	0.35	1.03	0.002	1.11

Table 8
Results of the Regression Analyses.

	R^2	*B*	*SE B*	*β*	*p*
Model 1: DV: Reflective Depth Post (*N* = 112)					
Step 1	.36				
Absolute Term/Constant		2.38	0.11		< .001
Reflective Depth pre		0.30	0.09	0.26	< .002
Control Group (yes = 0, no = 1)		-0.54	0.12	- 0.25	< .001
Contrast 1 (CG = 0, PG = 1)					= .76
Contrast 2 (CG = 0, *TLLS* = 1)		0.95	0.15	0.61	< .001
Contrast 3 (CG= 0, *TLLS*+external = 1)		0.80	0.18	0.40	< .001
Contrast 4 (CG = 0, *TLLS*+personal = 1)		0.73	0.19	2.34	< .001

Table 9
Analysis of the Significant Mean Differences for Increase of Reflective Depth (Reflective Breadth. 4 Levels) Based on the Intervention Type.

Intervention 1	Intervention 2	M1	M2	*p*	*d*[9]
CG	*TLLS*	0.19	0.77	< .001	1.11
CG	*TLLS* +External Video	0.19	0.78	< .02	0.61
CG	*TLLS* +Personal Video	0.19	1.07	< .002	0.82
PG	*TLLS*	0.55	0.77	< .001	0.85
PG	*TLLS*+External Video	0.55	0.78	< .008	0.78
PG	*TLLS*+Personal Video	0.55	1.07	< .001	0.83

[8] Only for the comparison of CG vs. *TLLS* a classical independent t-test was calculated, whereas all other comparisons were realized via U-test. The classical effect size *r* was calculated referring to Field (2012, p. 665) and converted into Cohens *d*.

[9] Only for the comparison of CG vs. *TLLS* a classical independent t-test was calculated, whereas all other comparisons were realized via U-test. The classical effect size *r* was calculated referring to Field (2012, p. 665) and converted into Cohens *d*.

Discussion Study II: Reflective Depth Development

It was found in this study that the reflective depth of English and physics teacher trainees differs measurably between participants of the five intervention types (regular *TLLS*, parallel *TLLS*, control group, and *TLLS* with personal/external video reflection), in consistent favor of the *TLLS* participants, and even with mostly high outcomes. Thus, the first part of H1 can be verified. However, we did not find evidence to support the hypothesis that with an additional video-supported reflection, reflective depth could be fostered to a stronger degree because non-significant differences were discovered. Thus, it seems safe to assume that reflective depth development occurs in the *TLLS* regardless of additional video reflections, with a gain of reflective depth of about one level rank (see *M*2, Table 8). Along with this is the fact that in the PG we do not see a comparable development (see *M*1 = 0.55, Table 8) despite a similar conceptual framework like the *TLLS.* This finding might serve to underpin the exclusive impact of the micro-interventions on the promotion of reflective depth.

As auspicious and positive as these results are, this study's findings do not resound in most of the other studies conducted on reflective depth development. Hatton and Smith (1995), for instance, have primarily detected the lowest mode of *descriptive reflective writing*, and discovered no indication of a fourth, critical level of reflection. Comparable results were delivered by Stender (2015), who could mostly extract *descriptive reflection* from the data material, only few *dialogic* and no *critical reflection* at all. Lüsebrink and Grimminger (2014, p. 208) also found no evidence to indicate pre-post-differences for teacher trainees' reflective depth. It remains unclear how far teacher trainees of the aforementioned studies were exposed to interventions to foster their reflective skills in such a way as it occurred during the *TLLS*. Contrary to the overwhelming corpus of such studies in which no effects were detected, reflective depth was promoted successfully in some other studies (Fund, Court, & Kramarski, 2002; Leonhard, Wüst, & Helmstädter, 2011). Eventually, the *TLLS*-participants demonstrate rather evaluative, analytic, and multi-perspective reflections, aspects associated with a proactive and learner-supportive stance and considered an attribute of expert teachers (Sato, Akita, & Naoki, 1993, p. 10). Novices' reflections on teaching were often found to be descriptive in style (Sabers, Cushing, & Berliner, 1991; Wolff, van den Bogert, Boshuizen, & Jarodzka, 2015, p. 80). These findings may imply that our *TLLS*-format might play a crucial role in modern teacher training, whilst fostering the reflective skills of the participants in depth and breadth.

The main limitations are that this study was conducted under quasi-experimental conditions, albeit with covariate control. Thus, teacher trainees could not be assigned to the *TLLS* in a randomized fashion even though it was later statistically secured whether participants' individual properties had affected reflective skill developments. Further, some of the estimates are imperfect due to small sample sizes and experimental mortality in the research process in some of the sub-groups, in particular the PG. Due to the complexity of the construct under investigation, not only one instrument should have been employed to elicit the reflective data to prevent mono-method bias. Besides that, prognostic and economic validity of this study are confined as we do not yet know how teacher trainees with high, medium, or low reflective skills will eventually behave in an authentic classroom setting in the foreseeable future.

References

Abels, S. (2011). *LehrerInnen als "Reflective Practitioner."* VS Verlag für Sozialwissenschaften.

Abendroth-Timmer, D. (2017). Reflexive lehrerbildung und lehrerforschung in der fremdsprachendidaktik: Ein modell zur definition und rahmung von reflexion. *Zeitschrift für Fremdsprachenforschung*, *28*(1), 101–126.

Abendroth-Timmer, D., & Frevel, C. (2013). Analyse handlungsleitender kognitionen anhand videogestützter reflexionsprozesse angehender spanischlehrender in verschiedenen berufsbiographischen kontexten. In U. Riegel & M. Klaas (Eds.), *Videobasierte Kompetenzforschung in Den Fachdidaktiken* (pp. 133–149). Münster, DE: Waxmann.

Akbari, R. (2007). Reflections on reflection: A critical appraisal of reflective practices in L2 teacher education. *System, 35*(2), 192–207.

Akbari, R. & Tajik L. (2009). L2 teachers' pedagogic knowledge base: A comparison between experienced and less experienced practitioners. *Australian Journal of Teacher Education, 34*(6), 52-73.

Allen, P. J. & Bennett, K. (2007). *SPSS for the health and behavioural sciences*. South Melbourne, AU: Thomson Learning.

Bandura, A. (1997). *Self-Efficacy: The exercise of control*. New York, NY: Freeman.

Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*(3), 200-212.

Birmingham, C. (2004). Phronesis: A model for pedagogical reflection. *Journal of Teacher Education 55*, 313-324.

Borko, H. & Livingston, C. C. (1989). Expert-Novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of Teacher Education, 40*(4), 36-42.

Bortz, J. (2010). *Statistik für human- Und sozialwissenschaftler*. Berlin, DE: Springer.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bullough, R. V. (1989): Teacher education and teacher reflectivity. *Journal of Teacher Education, 40*(2), 15-21.

Clarà, M. (2015). What Is reflection? Looking for clarity in an ambiguous notion. *Journal of Teacher Education, 66*(3), 261-271.

Combe, A. & Kolbe, F.-U. (2008). Lehrerprofessionalität: Wissen, Können, Handeln. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (pp. 857-875). Wiesbaden, DE: VS Verlag für Sozialwissenschaften.

Darling-Hammond, L., & Baratz-Snowden, J. (2007). A good teacher in every classroom: Preparing the highly qualified teachers our children deserve. *Educational Horizons, 85*(2), 111-132.

Davis, E. A. (2006). Characterizing productive reflection among preservice elementary teachers: Seeing what matters. *Teaching and Teacher Education, 22*(3), 281-301.

Day, C. (1993). Reflection: A necessary but not sufficient condition for professional development. *British Educational Research Journal, 19*(1), 83-93.

Dewey, J. (1933): *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston, MA: Heath.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160.

Fendler, L. (2003). Teacher reflection in a hall of mirrors: Historical influences and political reverberations. *Educational Researcher, 32*(3), 16-25.

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Sage Publications.

Freeman, D. & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly, 32*(3), 397-417.

Fund, Z., Court, D. & Kramarski, B. (2002). Construction and application of an evaluative tool to assess reflection in teacher-training courses. *Assessment & Evaluation in Higher Education, 27*(6), 485-499.

Gatbonton, E. (1999). Investigating experienced ESL teachers' pedagogical knowledge. *The Modern Language Journal, 83*(1), 35-50.

Gelfuso, A. & Dennis, D. V. (2014). Getting reflection off the page: The challenges of developing support structures for pre-service teacher reflection. *Teaching and Teacher Education, 38*, 1–11.

Griffiths, V. (2000). The reflective dimension in teacher education. *International Journal of Educational Research, 33*(5), 539-555.

Helbig, K., Günther, S. L., Rehfeldt, D., & Krüger, D. (2018). Umgang mit Schülervorstellungen zum Blutkreislauf: Validierung eines Videovignettentests. In M. Hammann & M. Lindner (Eds.), *Biologiedidaktik als Wissenschaft* (pp. 267–283). Dannstadt-Schauernheim: Aulis Verlag Deubner.

Häcker, T. (2017). Grundlagen und Implikationen der Forderung von Reflexivität in der Lehrerinnen- und Lehrerbildung. In T. Häcker & T. Leonhard (Eds.), *Reflexive Lehrerbildung Revisited* (pp. 21–45). Bad Heilbrunn: Klinkhardt.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81-112.

Hatton, N. & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49.

Klempin, C. (2019). *Reflexionskompetenz von Englischlehramtsstudierenden im Lehr-Lern-Labor-Seminar: Eine Interventionsstudie zur Förderung und Messung*. Stuttgart: J.B. Metzler.

Klempin, C., Rehfeldt, D., Seibert, D., Mehrtens, T., Nordmeier, V., Lücke, M., Köster, H., & Sambanis, M. (2019). *Realizing theory-practice transfer in German teacher education.* Retrieved from https://www.ristal.org/volumes/2019/volume-22019/news/realizing-theory-practice-transfer-in-german-teacher-education/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail&cHash=9f880294fe842858141ad8a3da8d642c

Jansing, B., Haudeck, H., Keßler, J.-U., Nold, G. & Stancel-Piatak, A. (2013). Professionelles Wissen im Studienverlauf: Lehramt Englisch. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Keßler & K. Schwippert (Eds.), *Professionelle Kompetenzen im Studienverlauf* (pp. 77–106). Münster: Waxmann.

Kaasila, R. & Lauriala, A. (2012). How do pre-service teachers' reflective processes differ in relation to different contexts? *European Journal of Teacher Education, 35*(1), 77–89.

Kirchhoff, P. (2017). FALKO-E: Fachspezifisches Professionelles Wissen von Englischlehrkräften: Entwicklung Und Valisierung Eines Domänenspezifischen Testinstruments'. In S. Krauss, A. Lidl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, R. H. Mulder (Eds.), *FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik* (pp. 113–52). Münster, DE: Waxmann.

König, J., Lammerding, S., Nold, G., Rohde, A., Strauß, S., & Tachtsoglu, S. (2016). Teachers'

professional knowledge for teaching english as a foreign language: Assessing the outcomes of teacher education. *Journal of Teacher Education*, *67*(4), 1–18.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology* (2nd ed.). London, UK: SAGE Publications.

Kuckartz, U. (2014). *Mixed methods: Methodologie, forschungsdesigns und analyseverfahren*. Wiesbaden, DE: Springer.

Kuckartz, U. (2016). *Qualitative inhaltsanalyse: Methoden, praxis, computerunterstützung*. Weinheim, DE: Beltz Juventa.

Legutke, M. K. & Schart, M. (Eds.) (2016). *Fremdsprachendidaktische Professionsforschung: Brennpunkt Lehrerbildung*. Tübingen: Narr Francke Attempto.

Leonhard, T., Wüst, Y., & Helmstädter, S. (2011). *Evaluations- und Forschungsbericht Schulpraktische Studien 2008-2010*. Heidelberg: Pädagogische Hochschule Heidelberg.

Lüsebrink, I., & Grimminger, E. (2014). Fallorientierte Lehrer/innenausbildung evaluieren: Überlegungen zur Modellierung von unterrichtsbezogener Reflexionskompetenz. In I. Pieper, P. Frei, K. Hauenschild & B. Schmidt-Thieme (Eds.), *Was der Fall ist: Fallarbeit in Bildungsforschung, Lehrerbildung und frühpädagogischen Ausbildungs- und Berufsfeldern* (pp. 201–212). Wiesbaden: VS Springer.

Neuweg, G. H. (2007). Wie Grau Ist Alle Theorie, Wie Grün des Lebens goldner Baum? *LehrerInnenbildung Im Spannungsfeld von Theorie Und Praxis*, *12*, 1-14.

Rehfeldt, D., Seibert, D., Klempin, C., Mehrtens, T., Lücke, M., Sambanis, M., Köster, H., & Nordmeier, V. (2018). Mythos Praxis um jeden Preis? Die Wurzeln und Modellierung des Lehr-Lern-Labors. *Die Hochschullehre*, *4*, 90–114.

Rehm, M. & Bölsterli, K. (2014). Entwicklung von Unterrichtsvignetten. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (pp. 213–25). Berlin; Heidelberg: Springer Spektrum.

Rieger, A., Radcliffe, B. J., & Doepker, G. M. (2013): Practices for developing reflective thinking skills among teachers. *Kappa Delta Pi Record, 49*(4), 184–189.

Rodgers, C. (2002). Seeing student learning: Teacher change and the role of reflection. *Harvard Educational Review, 72*(2), 230–253.

Roters, B. (2012). *Professionalisierung durch Reflexion in der Lehrerbildung: Eine empirische Studie an einer deutschen und einer US-amerikanischen Universität*. Münster; München: Waxmann.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*(1), 63–88.

Santagata, R., Zannoni, C., & Stigler, J. W. (2007). The role of lesson analysis in pre-service teacher education: An empirical investigation of teacher learning from a virtual video-based field experience. Journal of Mathematics Teacher Education, 10(2), 123–140. https://doi.org/10.1007/s10857-007-9029-9

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education, 61*(4), 339–349.

Santagata, R., & Yeh, C. (2013). Learning to teach mathematics and to analyze teaching effectiveness: Evidence from a video- and practice-based pre-service course. *Journal of Mathematics Teacher Education, 17*, 491-514.

Sato, M., Akita K., & Iwakawa, N. (1993). Practical thinking styles of teachers a comparative study of expert and novice thought processes and its implications for rethinking teacher education in Japan. *Peabody Journal of Education, 68*, 100–110.

Schädlich, B. (2015). Fachpraktika im Master of Education Französisch aus der Perspektive der Studierenden: Ein Beitrag zur Entwicklung reflexiver Handlungskompetenz? *Zeitschrift für Fremdsprachenforschung, 26*(2), 255–285.

Schmitz, G. S. (2000). *Zur Struktur und Dynamik der Selbstwirksamkeitserwartung von Lehrern. Ein protektiver Faktor gegen Belastung und Burnout?* Berlin: Freie Universität Berlin.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco, CA: Jossey-Bass.

Sherin, M. G. & Van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education, 60*(1), 20-37.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Smyth, J. (1989). Developing and sustaining critical reflection in teacher education. *Journal of Teacher Education, 40*(2), 2–9.

Steinke, I. (2007). Qualitätssicherung in der qualitativen Forschung. In U. Kuckartz, H. Grunenberg, & T. Dresing (Eds.), *Qualitative Datenanalyse: computergestützt: Methodische Hintergründe und Beispiele aus der Forschungspraxis* (pp. 176–187). Wiesbaden: VS Verlag für Sozialwissenschaften.

Stender, J. (2015). *Die Bedeutung von Reflexionskompetenz im Zuge der Professionalisierung von angehenden*

Lehrpersonen. Retrieved from http://nbn-resolving.de/urn:nbn:de:hbz:466:2-24552

Tschannen-Moran, M., Woolfolk Hoy A. & Hoy, W. K (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*(2), 202–248.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1-67.

Wolff, C. E., van den Bogert, N., Boshuizen, H. P. A., & Jarodzka, H. (2015). Keeping an eye on learning. Differences between expert and novice teachers' representations of classroom management events. *Journal of Teacher Education, 66*(1), 68–85.

Wildt, J. (1995): Feedback in Lehrveranstaltungen. In W.-D. Webler & J. Osterloh (Eds.), *Gute Lehre Konzepte, Methoden, Nachwuchsförderung* (pp. 43-51). Weinheim: Beltz.

DR. CHRISTIANE KLEMPIN is a postdoctoral research assistant at the Dahlem School of Education at Freie Universität Berlin. Her field of expertise is the assessment of the effectiveness of English language teacher education and its evidence-based optimization. From 2015 until November 2019, she adapted and investigated so-called Teaching and Learning Labs (TLL) for the specific needs of English student teachers. Her research on *TLL* yielded insights into its positive effects on teacher trainees' reflective skills. Her second project (2019-2022) focuses on the theories of Performative Didactics and how performative instruction and attitudes of trainees attending a *TLL* can be measurably honed.

DR. DANIEL REHFELDT concluded his PhD in 2017 in the Department of Physics Teacher Education. As part of his dissertation, he developed a quantitative measurement instrument to assess the quality of experimental laboratory practicums during students' higher science training. His PhD was followed by a postdoctoral research phase from 2017 until 2018 aiming to diversify the above-mentioned *TLL*-teacher training format to other subject domains than physics (e.g. English, history, and primary science education). Since 2018, he holds a position as a permanent postdoctoral researcher in primary science education.