What Makes Observational Learning in Teacher Education Effective?

Evidence from a meta-analysis and an experimental study

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

Vorgelegt von

Olga Chernikova

München, den
02.05.2018

1st Supervisor / Erstgutachter:	PD Dr. Karsten Stegmann
2nd Supervisor / Zweitgutachter:	Prof. Dr. Jan-Willem Strijbos
3nd Supervisor /Betreuerin:	Prof. Dr. Tina Seidel
Tag der mündlichen Prüfung:	16.07.2018

Acknowledgements

I would like to sincerely thank all the people, who supported and inspired me on my way of conducting research and completing this doctoral dissertation.

First, I would like to express my deep gratitude to my supervisors, PD Dr. Karsten Stegmann and Prof. Dr. Jan-Willem Strijbos for their highly professional guidance and patience, for encouraging me to explore and keep up with the recent research practices and methodological approaches, and for the useful critical feedback during all stages of planning, conducting research, and completing this dissertation. I would also like to thank Prof. Dr. Tina Seidel for her valuable advice and assistance during planning and preparing the empirical part of my dissertation.

Second, I would like to thank Dr. Phil. Thomas Froschmeier from the Department of Sports and Health Sciences at Technical University of Munich for cooperation and support in planning and conducting data collection during Basic Qualification Sports course in 2016 and 2017. I would also like to appreciate the work of Nikolaus Hawranek and all the coordinators and the teaching team of the course for readiness to help and support my research and especially for the language support provided. My gratitude also extends to all the participants of the course for their time and effort.

Third, special thanks should be given to Georg Kuschel for technical support, providing equipment and patient instructions for data collection processes. I would also like to thank my colleagues from the department for fruitful discussions during conferences and retreats and great emotional support in times of difficulties.

Last, but not least, I would like to thank my family and friends for believing in me, being patient and providing all possible care and support for me to complete the work on this doctoral project. Special appreciation to Oleg and Julia Khojanov for the inspiration and the shelter, where the most of the dissertation was written. Thank you for the support!

Contents

List of Tables

List of Figures

Summary

This doctoral dissertation is completed in the area of Teacher education and aims at researching how observational learning can be implemented to effectively foster teaching competence.

The Chapter 2 introduces a meta-analysis on the effects of the observational learning on acquisition of teaching skills. In this meta-analysis, observational learning was defined as observing the target skill modelled by an expert and adopting this skill in one's own practices. The main goal of the meta-analysis was to assess the effectiveness of observational learning in teacher education and the moderating role of the (1) presentation format used to present the target skill, (2) scaffolding and (3) outcome measures in fostering and assessing knowledge and skill acquisition The research questions were:

(1) To what extent does the observational learning in Teacher Education affect objective and subjective learning outcomes?

(2) To what extent does a presentation format and measures of performance influence the effectiveness of observational learning?

(3) To what does scaffolding influence the effect of the observational learning on learning outcomes?

The meta-analysis summarised 19 independent empirical research findings between 1969 and 2014 based on the procedure suggested by Borenstein and colleagues (2009). Furthermore, the role of several methodological issues in relation to research in the domain of teacher education were addressed (i.e., using a quasi-experimental design, using relatively small samples, not having pure control conditions, etc.) and multiple statistical methods were combined to ensure the quality and validity of the results and to control for possible publication bias and questionable research practices. The findings went in line with the Bandura´s (1986) and Chi´s (2009) theoretical framework of observational learning and also supported the

assumption that instructionally supported observational learning is beneficial for the acquisition of complex skills in the domain of teacher education. The meta-analysis came across several limitations and raised additional questions that were partially answered by the empirical study presented in Chapter 4.

The Chapter 3 introduces the development of an instrument. More specifically, establishing and validating a scale to measure lesson planning competency of pre-service teachers (as a part of teaching competence in general). The target user group were pre-service elementary school teachers in the domain of physical education, but, with slight modifications, the procedure and conceptual considerations behind the scale can also be used for different domains of teaching, as well as for teachers with differing levels of expertise. The main goal of the scale development chapter was to close the gap between (1) assessing local effects on the acquisition of very specific teaching skills and (2) assessing a level of teaching competence as a more general construct, which combines different types of knowledge and skill. The research goals were:

(1) To test if lesson planning competency can be measured as a single construct with different processes having different difficulty (noticing on the easier side of the scale, analysing and explaining in the middle and suggesting new ideas on a more difficult part of the scale), but building upon each other to define lesson planning competency.

(2) To select appropriate materials and develop a scale representing the complex skill of lesson planning competency. It was assumed that items of the scale could be clustered and that the specific competency level could be assessed based on item difficulty.

The created scale met the assumptions of the Item Response Theory, achieved accepted level of reliability (above .65) and was used in empirical study, presented in Chapter 4.

The Chapter 4 describes an empirical study aimed at assessing the effect of observational learning and scaffolding on fostering the lesson planning competency in elementary school pre-service physical education teachers. The main assumption was that

observational learning would be beneficial, to foster lesson planning competency and thereby contribute to the development of teaching competence. The research questions of the empirical study were:

(1) To what extend does scaffolding (facilitating the formulation of learning goals) during observational learning, impacts the pre-service teachers´ lesson planning competency?

(2) To what extend do teaching experience and motivational factors (beliefs about the importance of learning goals) predict the post-test lesson planning competency?

(3) To what extend does adherence to instructions during the treatment phase predict the lesson planning competency at the post test phase?

Although no significant differences were identified between treatment and control condition during the post-test, the study supported the hypothesis, that following the scaffolding procedure suggested to experimental condition had a positive effect on the lesson planning competency level, and in general supported the assumption that observational learning can be used for fostering this competency.

1. General Introduction

This chapter serves as an introductory part of doctoral dissertation and provides information about the state of the art research in Teacher Education concerning teaching competence and use of observational learning as a teaching and learning strategy to foster the development of the teaching skills. It also provides an overview of the dissertation´s structure.

1.1 Problem statement

Learning new skills and behavioral patterns from others through observation is one of the most common ways to learn. According to Bandura (1986), observing others is one step towards learning the observed skill, as the observation fosters (1) the initial steps of creating a sort of a cognitive schema of how and when the skill is applied, an *internal script* of the skill (cf. Fischer et al., 2013), (2) affects motivation by providing information on the success of the skill (cf. Bandura, 1986) and (3) results in – compared to problem-based learning or learning-by-doing – lower extraneous cognitive load (cf. Renkl & Atkinson, 2003; Sweller, 2005). The existence of a basic internal script of a skill can be regarded as a prerequisite to practice a skill and further develop the both understanding the skill as well as its performance. Therefore, observational learning is a promising learning and teaching strategy that can be used in different domains (Chi, Hausmann & Roy, 2008; Hoover, Giambatista & Belkin, 2012; Stark, Kopp & Fischer, 2011; Stegmann, Pilz, Siebeck & Fischer, 2012).

While there is a strong body of evidence that learning from observations (like worked examples) has a substantial medium positive effect on (cognitive) skill acquisition in general (Crissmann, 2006), observational learning of social interaction or similar skills including are comparatively rare in many domains. An exception is the learning of communication skills in the field Medical Education (e.g., Stark, Kopp & Fischer, 2011; Stegmann, Pilz, Siebeck & Fischer, 2012). The findings from Medical Education show that observational learning can effectively facilitate complex skill acquisition (Heitzmann, Fischer, Kühne, Eversmann & Fischer, 2015). Another area in which the effect of observational learning is examined is

Teacher Education. Much of what pre-service and starting in-service teachers need to be aware of and implement in practice cannot be learned solely through context-independent instruction, but rather requires exposure to the authentic contexts that the teachers will later encounter in practice. Therefore, classroom observation presents an opportunity to see, experience and evaluate real-life teaching situations, including responses to difficult classroom situations as modelled by experienced teachers.

On the long run, observational learning can enable pre-service teachers to become adaptive experts, that is, to make judgments in the face of uncertainty, to innovate, and to be able to continuously learn from their practice (Darling-Hammond, Hammerness, Grossman, Rust, & Shulman, 2005). Observational learning is a tool, that can prepare pre-service teachers to learn from practice, by offering essential strategies/skills to analyse observed behaviors and improve own teaching (Santagata, Zannoni & Stigler, 2007). It allows learning in an authentic but safe environment, in which pre-service teachers face situations and challenges similar to the ones they will experience in real classrooms. Observational practices are sporadically used to foster acquisition of the range of teaching skills, but also to assess and provide feedback on the teachers´ performance. Because observation is used in different contexts and with different purposes, there is a lack of systematic knowledge of observational learning, as a teaching and learning strategy in Teacher Education.

Despite the obvious strengths of observational learning, research has shown that several factors can lead to suboptimal learning processes and outcomes (Stegmann, Pilz, Siebeck, & Fischer, 2012): the complexity of the observed situation, leading to increased cognitive load and not being able to notice/process important details of the learning situation; focus on superficial characteristics of the learning situation (i.e room settings, teachers´ manner of speech or appearance, emotional reactions of students or teacher), rather than on its core elements (i.e. learning and teaching concepts and strategies, application specific techniques to achieve learning goals); not being able to connect observed situation or its elements to

conceptual theoretical knowledge and make generalizations. These factors characterize an unstructured observation and limit the effectiveness of learning by observation. Different types of scaffolding and/or additional instruction seem to be able to overcome the limitations of an unstructured observational learning, by focusing the attention on the core elements of the behavior to be learned, knowledge, skills or competency (Chi et al., 2008; Dianovsky & Wink, 2011; Glogger et al., 2009; Hübner, 2009; Van Gog & Rummel, 2010). Up to this date, there is no systematic review of observational learning in Teacher Education that would consider design features of observational learning (presentation of a target skill, the amount and type of scaffolding provided, the tasks designed to measure the learning outcomes) and their effect on the teachers´ learning.

Over the past several decades, the professional competence of teachers was studied from several perspectives. Some studies focused on defining the teaching competence and its components (Epstein & Hundert, 2002; Koeppen, Hartig, Klieme, & Leutner, 2008), and performed explorative and descriptive studies. Other studies introduced interventions to foster specific teacher skills (e.g., classroom management, presentation, setting up working environment, etc.), based on a single didactic principle (Crooks & Gifford, 1992; Koran Jr., 1969; Koran Jr., 1970; Slogget, 1972). Although the literature often uses the term “teacher competence”, this dissertation uses the term “*teaching* competence” to emphasize that the focus is on the activities performed by a teacher, rather than various stable teachers´ characteristics. This approach is in line with recent findings by Hiebert and Stigler (2017), who argue that improving teaching as a system is more promising than focusing on the teacher as a single element of this system and on improving the teacher’s characteristics. As teaching competence is not merely the addition of separate skills, but rather a broader construct, which involves a combination of knowledge, skills and attitudes (Blömeke et al., 2015), measures of acquisition of a single didactic principle can hardly provide enough information to assess the level of teaching competence or provide insights for its structure, development and

improvement. In light of this fact, there is also a lack of empirical research on teaching competence as a complex construct. More research is needed to specify the core components of the competence, specific teaching task characteristics, and, subsequently, develop a measurement instrument to assess competence as a complex construct and assess effectiveness of teaching and learning strategies aimed at fostering teaching competence.

Teaching competence is a broad construct, which involves a combination of knowledge, skills and attitudes (Blömeke et al., 2015). The term ‘‘Competence’’ (plural ‘‘competences’’) is the broader term and is used in holistic approaches; the term ‘‘competency’’ (plural ‘‘competencies’’) is used in analytic approaches, is considered to be a part of competence and focuses rather on task characteristics and the elements of the tasks to be performed effectively (Stoof et al., 2002). One of the essential tasks a teacher has to perform is planning the lesson (Duplass, 2006; Jensen, 2001). Lesson planning helps to produce a unified structure of the lesson (Jensen, 2001), which in turn gives teachers the opportunity to deliberately think about and set the learning goals, select teaching and activities, and materials needed. Both competence and competency are regarded as learnable and have thus the potential to be improved (Epstein & Hundert, 2002; Shavelson, 2010; Weinert, 2001). It is important to notice, that both are also domain specific (Blömeke et al., 2015) which means that for a teacher to perform efficiently, s/he needs not only general pedagogical, but also domain specific knowledge and skills.

1.2 Aims of the Dissertation and Overarching Research Questions

This dissertation aims at contributing to the theoretical and empirical body of research in use of observational learning in teacher education by addressing the following issues: (1) conducting a systematic review and meta-analysis on the effects of observational learning on teaching related skills and teaching competence in general as well as the role of design related features (presentation format of the target skill, use of scaffolding and additional instructional support, use of different measures to assess learning outcomes); (2) designing and conducting

an empirical study to address the use of the observational learning to foster the development of (the part of) teaching competence, namely lesson planning competency. The empirical study is conducted in the field of physical education, which, on the one hand, provides broad opportunities to use observational learning in developing pre-service teachers´ competence, and, on the other hand, has not been sufficiently researched so far. A specific scale was developed to assess the lesson planning competency as a complex construct, rather than a single didactic skill. The scale aimed at addressing the complexity of the competency and being sensitive enough to measure changes in competency acquisition. Therefore, this dissertation will also contribute to the practical side of research in teaching education by addressing methodological aspects of measuring teaching competence.

The overarching research questions of this dissertation are to identify (1) if observational learning is an effective teaching/learning strategy that contribute to fostering the pre-service teachers´ competence; and (2) in what way should observational learning be organised and designed to ensure that the target competency is acquired in the most effective way. To answer these questions, first, the systematic review and the meta-analysis of the effects of observational learning in Teacher Education and the role of such design features as a presentation format of a target skill, measures used to assess learning and different types of scaffolding to support observation was conducted. The domain of physical education, was identified as one of the areas where empirical evidence of effective learning is lacking and was used for the current empirical study. Second, a scale was developed to assess a part of teaching competence (a lesson planning competency) as a complex construct to provide instrumental support and address the research questions of the planned empirical study. Third, after creating and validating the scale, the experimental study was conducted, where scaffolding was implemented to support initial stage of lesson planning (the goal formulation).

1.3 The structure of doctoral dissertation

The structure of the dissertation is demonstrated in Figure 1.1. The dissertation consists of three main chapters: (1) a meta-analysis on the effects of observational learning on learning outcomes in teacher education; (2) the development of a scale to measure the lesson planning competency in physical education pre-service teachers; and (3) an empirical study on the role of scaffolded observational learning in fostering lesson planning competency of physical education elementary school pre-service teachers.

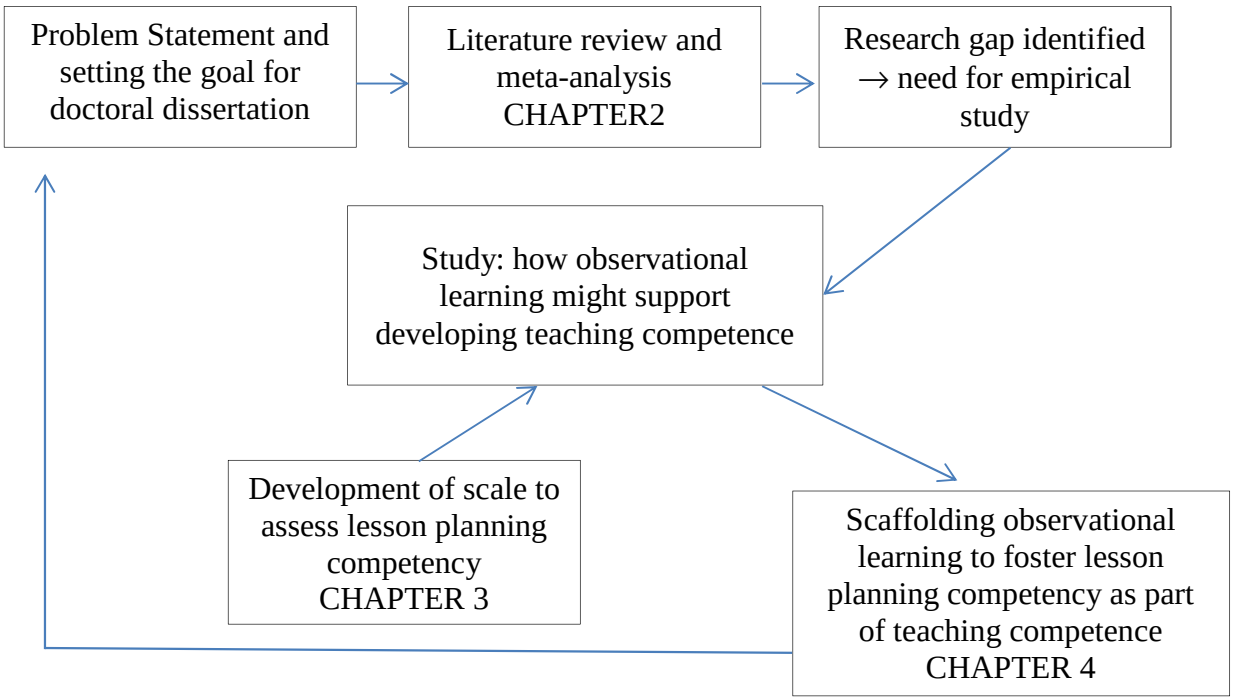


Figure 1.1. Structure of doctoral dissertation.

As no systematic review of the effects of observational learning on learning outcomes in Teacher Education had been conducted before, this became a starting point for this dissertation. The Chapter 2 presents a meta-analysis on the effects of observational learning on the acquisition of teaching skills. In this meta-analysis observational learning is defined as observing the demonstration of a target skill modelled by an expert and adopting the skill in one's own practices. The main goal of the meta-analysis was to assess the effectiveness of the observational learning in teacher education and the role of the presentation format used to

present a target skill, scaffolding and outcome measures in fostering and assessing knowledge and skill acquisition The research questions set for the meta-analysis were: (1) to what extent does the observational learning in Teacher Education affect objective and subjective learning outcomes? (2) To what extent does a presentation format (in vivo, video-based or text-based) and measures of performance (actual performance or text-based skill test) influence the effectiveness of observational learning? (3) To what does scaffolding moderate the effect of the observational learning on learning outcomes?

The meta-analysis summarised 19 independent empirical research findings between 1969 and 2014 based on the procedure suggested by Borenstein and colleagues (2009). Furthermore, the role of several methodological issues in relation to research in the domain of teacher education were addressed (i.e., using a quasi-experimental design, using relatively small samples, not having pure control conditions, etc.) and multiple statistical methods were combined to ensure the quality and validity of the results and to control for possible publication bias and questionable research practices.

The Chapter 3 details the development of an instrument. More specifically, establishing and validating a scale to measure lesson planning competency of pre-service teachers (as a part of teaching competence in general). The scale was developed and validated on a group of pre-service elementary school teachers in the domain of physical education. Item response theory was applied to resolve the methodological issues in measuring lesson planning competency. The main goal of Chapter 3 was to close the gap between assessing local effects in terms of the acquisition of specific teaching skills and assessing a level of teaching competence as a more general measure of knowledge and skill acquisition. The specific research goals set for the instrument development were (1) to test if lesson planning competency can be measured as a single construct with different processes having different difficulty (noticing on the easier side of the scale, analysing and explaining in the middle and suggesting new ideas on the more difficult part of the scale), but building upon each other to

define lesson planning competency; (2) to select appropriate materials and develop a scale representing the complex skill of lesson planning competency, assuming that the items of the scale can be clustered and that the specific competency level can be assessed based on item difficulty.

The Chapter 4 describes an empirical study aimed at assessing the effect of observational learning and scaffolding on fostering the lesson planning competency in elementary school pre-service physical education teachers. The main assumption was that observational learning will be beneficial to foster lesson planning competency and thereby contribute to the development of teaching competence. The research questions for the empirical study were: (1) to what extend does scaffolding (facilitating the formulation of learning goals) during observational learning, impacts the pre-service teachers´ lesson planning competency? (2) To what extend do teaching experience and motivational factors (beliefs about the importance of learning goals) predict the post-test lesson planning competency? (3) To what extend does adherence to instructions during the treatment phase predict the lesson planning competency at the post-test phase?

The Chapter 5 sums up and discusses findings from Chapters 2, 3 and 4, encountered conceptual and methodological issues and limitations in the preceding chapters, as well as the theoretical and practical implications, and the directions and insights for further research in teacher education.

2. A Meta-Analysis on the Effects of Observational Learning in Teacher Education

Showing learners procedures step-by-step is a common approach to teach specific skills, e.g. a teacher models division in Math classes, a sports teacher demonstrating a course of motions, a teacher modeling translation of sentences in Latin classes. According to Bandura (2001), observing others is one step towards learning the observed skill. Actively observing others performing a skill is assumed to (1) support the creation of some basic, rudimentary understanding and a kind of internal script of the skill (cf. Fischer et al., 2013), to (2) affect motivation by providing information on the success of the skill (cf. Bandura, 1986) and (3) to imply – lower extraneous cognitive load compared to problem-based learning or learning-by-doing approaches – (cf. Sweller, 2005). The existence of a rudimentary internal script of a skill can be regarded as prerequisite to practice a skill, todevelop the internal script of performing the skill as well as the actual performance of the skill. The knowledge that a specific procedure allows to solve problems which previously (subjectively) seemed to be unsolvable may increase the motivation of learners to learn and apply a new skill. Moreover, reduction of extraneous cognitive load allows for more capacity to create or develop internal scripts for new skills to be learned and integrated with previous knowledge and experience.

While there is a strong body of evidence that learning from observations (i.e. using worked examples) has a general positive effect on (cognitive) skill acquisition (Crissmann, 2006), using an observation to learn the skills that include social interaction is less common. One exception is learning communication skills in the field of Medical Education (e.g., Stark, Kopp & Fischer, 2011; Stegmann, Pilz, Siebeck & Fischer, 2012). Another area of research where the effect of observational learning is examined is Teacher Education. The evidence coming from Medical Education claims that observational learning can facilitate complex skill acquisition effectively (Heitzmann et al., 2015).However, while studies in Medical Education provide rather strong empirical evidence (i.e. experimental designs, randomization, sufficient sample sizes) at the first glance, the evaluation of empirical evidence in studies on

observational learning in Teacher Education is more ambiguous: studies with relatively large samples often do not quantify their results (e.g., Beswick & Muir, 2013; Henninger, 2002; Wang, 2013), while studies with small sample sizes use rather complex designs in relation to their sample size (e.g., Koran et al, 1972; Kubany & Sloggett, 1991). Furthermore, several studies have quasi-experimental designs, which make causal attributions problematic (e.g., Claus, 1969; Crooks & Gifford, 1992).

Against this background, this meta-analysis aims to provide a systematic review of the quantitative empirical studies on the effects of observational learning (compared to traditional approaches) on skill acquisition in Teacher Education. The described methodological issues require a careful and multi-methodological approach to identify and – if identified – to correct for potential biases. Therefore, first the theoretical background regarding the observational learning, including potential moderating factors is introduced to prepare the research questions. Furthermore, methodological issues regarding specific features of empirical research on the effects of observational learning in Teacher Education were addressed. This is followed by an overview and a discussion of methodological approaches to test and to handle potential biases like publication bias or bias caused by questionable research practices, namely Egger's test (Sterne & Egger, 2001), Trim'n'fill (Duval & Tweedie, 2000), p-curve analysis (Simonsohn, Nelson, & Simmons, 2014), R-index (Schimmack, 2012), and fail-safe N (Rosenthal, 1979).

2.1 Observational Learning in Teacher Education

2.1.1 Cognitive and Socio-cognitive theories and mechanisms

The rationale of observational learning is based on Bandura's (1986) social cognitive theory, which explains learning as a continuous interaction between cognition, behavior and environment. The social cognitive theory stresses that observational learning relies strongly on the attention during observation, the memory after observation as well as the motivation to perform the skill and to actually perform the skill.

Since not all observed behaviors can be or are effectively learned, there is a need to identify the factors involving both the model and the learner that play a role in whether learning is successful and whether it leds to actual changes in the behavior. Among these factors Bandura (1986) named attention, retention, reproductive, and motivational processes.

Previous empirical research has also shown that an observation can be effective for learning in both well- and ill-structured domains. For example, an observation was found to have a positive effect in creative domains (Groenendijk, Janssen, Rijlaarsdam & Van den Bergh, 2013); text writing (Braaksma, Rijlaarsdam & Van den Bergh, 2002; Couzijn, 1999; Raedts, Rijlaarsdam, Van Waes & Daems, 2007); learning to collaborate and cooperate (Rummel & Spada, 2005; Schworm & Renkl (2007; Van Steendam, Rijlaarsdam, Sercu & Van den Bergh, 2010).

An observation is also an important part of teachers' education. Much of what pre-service and beginner teachers need to be aware of cannot be learned solely in a context-independent environment. Therefore, observations of real or modeled classroom situations provide an opportunity to see, experience and evaluate real-life teaching situations as well as responses to difficult classroom situations as they are modeled by experienced teachers. As modeling and the observational learning theories were evolving, efforts to relate those to teachers' education were made. Different instructional approaches of teacher education like microteaching, student teaching, performance assessments and portfolios, analyses of teaching and learning, case methods, autobiography, and practitioner inquiry were intended to support teachers' abilities to learn in and from practice (Allen & Ryan, 1969; Darling-Hammond et al., 2005; Santagata, Zannoni, & Stigler, 2007).

Novice teachers can observe the behavior modeled by expert teachers and in this way learn to apply conceptual knowledge into practical tasks as well as to learn specific skills for classroom management, teaching techniques, etc. Observing the behavior of more experienced colleagues and learning skills from them is what defines the observational learning (Chi, 2009).

The observational learning is supported by recent research as a method that enables novices to learn even complex cognitive skills through an observation (Chi, Roy & Hausmann, 2008; Fryling, Johnson & Hayes, 2011).

Although empirical research shows positive findings, the nature of the observational learning as described by Bandura (1986) presupposes some possible drawbacks. If the observational learning naturally occurs in social settings, not all outcomes related to it are advantageous. Learners observe the behavior demonstrated by a role model as well as the consequence of the behavior. This also implies that if role models demonstrate poor behaviors, this can also be learned. Furthermore, if the consequences of the poor behavior are not clear, undesirable models can reinforce that particular behavior. It it is therefore essential to have a good role model, demonstrating desirable behavior. Coming from naturally occurring to instructionally supported observational learning used in education, the problem of good role models can be regarded as less significant; another problem however increases in importance: observing the behavior and learning it does not necessarily lead to changes in the behavior (Bandura, 1986).

Even if the behavior is well modeled, what is noticed and what is adopted by learners, especially in complex domains, requires genuine concern. Motivation becomes an important factor. According to Bandura (1986) learners seem to be more motivated to repeat behaviors they enjoy and are capable of performing successfully. Individual differences and capacities, as well as proper instructional support (to attract students' attention to the essential elements) should be adopted to make the observational learning successful. Observation in this case is referred to as an active, purposeful task that stimulates deep learning and the development of professional knowledge and skills (Hanson, Bannister, Clark, & Raszka, 2010).

Hoover, Giambatista, & Belkin (2012) also mention some drawbacks that might be applied to learning solely by observation – cognitive bias may lead learners to screen out plausible alternatives essential to effective coding, but not being actually faced with

uncomfortable aspects of the situation. Additionally, as suggested by social cognitive theory (Bandura, 1986) this technique does not usually work effectively if observation is spontaneous and not structured or supported.

There is also evidence that learning complex skills through observation requires additional scaffolding either before, during, after observation or continuously (Chi et al., 2008; Dianovsky & Wink, 2011; Glogger et al., 2009; Hübner, 2009; Van Gog & Rummel, 2010). This scaffolding can come in different formats and intensity: focusing students on specific behavior during observation, introducing questions for discussion or cognitive prompts, giving guidelines for making notes, etc.

2.1.2 Skills that can be learned through observations in teacher education

Professional competency of teachers involves the ability to plan, understand and analyze classroom situations. It requires not only knowledge about concepts, theories and principles, but also the ability to apply abstract knowledge in the classroom in a way that meets both the formal requirements of educational systems and the individual needs of learners in the context of the current classroom situation.

Observing experienced teachers can, on the one hand, allow learning possible techniques to address diagnostic of learning difficulties in students, difficult situations in classroom, classroom management, attracting attention, raising students' motivation, etc. On the other hand domain specific approaches and techniques can be learned: such as using example-based and problem-based lessons, using questioning techniques, argumentation, etc. Observation can also help with analyzing different stages and elements of the lesson, notice learning goals, good and poor practices and their influence on classroom dynamics and student learning. This in turn can improve pre-service teachers' planning of own teaching activities and understanding of possible strengths and weaknesses of teaching approaches.

A recent literature review by Gaudin and Chalies (2015) assessed the role of video materials increasingly used in teacher education to address various skills and competencies of

future teachers. The authors used a categorization of knowledge and skills which can be learned through observing video, which also applies to other types of observation. Besides knowledge and skills selective attention, knowledge-based reasoning, building knowledge on "how to interpret and reflect", and on "what to do" were also mentioned. For the current study presenting common and the best practices (and acquiring "what to do" type of teachers' knowledge) became the core element of study.

2.1.3 Presentation format of the modelling in the observational learning

The idea to provide pre-service teachers with authentic models of classroom situations, possible problems and teacher behaviors was supported throughout the whole history of teacher education. The presentation format of the models to be observed varies a lot between studies. Observed can be real or simulated classrooms on video or in vivo (direct observation), text records and combinations of where visual and text cues are used.

Direct observation in the context of pre-service teachers' learning from observation refers to observing the real or simulated lessons by being physically present at observation site, but not participating in teaching process (Gettinger & Stoiber, 2014; Lavin, 1992). This method is widely used in pre-service teachers' field practice and in experimental research, when pre-service teachers are placed in or outside the classroom so that they can observe the lesson without altering the classroom environment.

With the development of the technology, video models became more popular (Gaudin & Chalies, 2015) as they need less resources and time to be prepared and at the same time allow focusing on specific behavior, and can be shown to relatively big audiences. Recorded videos can be re-watched at a later point; they can also be paused and discussed at any point. Therefore videos are now often used for educational and feedback purposes (Zottmann, et al., 2013) in and beyond teacher education.

Another possible presentation format is using a text script of the lesson or case-study describing teachers' behavior in written format. Moreno and Valdez (2007) report video cases

to be more effective than text narratives regarding motivation and transfer of knowledge, but both similarly effective when it comes to knowledge and skill acquisition. Furthermore, written scripts can be less distracting than the video of classroom situation or observing it in vivo, as they presents less superficial details of learning situation, but gives exactly the same information about leaning content. Written scrits also requires fewer costs.

Koran et al., 1971 reports that written lesson scripts are as just as effective for skill acquisition as videos are. On the one hand, they also allow for repeated exposure which in turn allows for more details to be noticed. On the other hand, in comparison to video material, text format is definitely perceived as less authentic and possibly less engaging. Every format has its own costs and restrictions, which in turn led to the question of what presentation format is the most effective in pre-service teacher education.

2.1.4 Scaffolding Skill Acquisition in Observational Learning

Complex cognitive skill acquisition, according to Van Lehn (1996), involves deliberate retrieval, mapping, and application, generalization of principles, as well as transfer of the principles to different tasks. This also applies to the range of skills essential for teachers. Professional competency of teachers involves the ability to plan, to understand and to analyze classroom situations through applying pedagogical and domain conceptual knowledge, classroom management principles and finishing with the ability to learn while teaching. Although some of the skills can be developed spontaneously, the general concern is that processing new information usually takes place in the working memory which is known to be limited in capacity. From this point of view, any instructional method that ignores this fact has been discussed as ineffective for learning (Krischner, Sweller & Clark, 2006). This concern emphasizes the role of the scaffolding in education in general, but also in observational learning as one of the educational methods.

Learning by observation can be efficient only if students are actively involved in the process (Bandura, 2001; Chi et al., 2008). To define active observation, the current study is

grounded on the so-called ICAP framework, a framework that differentiates student engagement in learning tasks by categorizing their behaviors as Interactive, Constructive, Active, or Passive (Chi, 2009; Menekse, Stump, Krause, & Chi, 2013). In relation to the current study, the ICAP framework can be considered as one of the ways to scaffold students *during* observation, suggesting that activating students' attention and participation in observation is more beneficial than just passive observation. This approach focuses on supporting students during skill acquisition by focusing their attention and decreasing distraction and confusion. Additional possible scaffolds are offered by discussions, cognitive prompts or questions, protocols, learning diaries, etc. These methods provide support primarily *after* observing the behavior and help to make observed behavior part of the personal/internal experience by putting experiences into words, structuring and discussing with others.

For example, previous research hypothesized that writing notes in learning diaries can support learning due to several reasons. Learning diaries are believed to provide an opportunity for students to discover their personal ways of learning and understanding of the course content (Glogger, et al., 2009). They also encourage learners to get involved in constructive behaviors, to generate connections between ideas, link together concepts and make sense of the overall picture (Dianovsky & Wink, 2011).

Although general research finding favor use of cognitive prompts for knowledge acquisition (Berthold, Nuckles & Renkl, 2007; Froschmeier, Stegmann, Zottmann & Matikalo-Siegl, 2012; Glogger et al., 2009; Hübner, 2009; Schworm & Renkl, 2007; Stegmann et al., 2012;), there are few problematic issues that need thorough consideration. A common matter of concern for additional instructional support is represented by the cognitive load perspective (Hoover et al., 2012; Sweller, Ayres & Kalyuga, 2011). Wrong application of instructional support can either lead to overload for students with lower prior knowledge or low self-assessment and decrease of motivation for learners with higher prior knowledge (Reisslein, Atkinson, Seeling & Reisslein, 2006). Research in cognitive load theory has also shown that in

the initial stage of skill acquisition, learning from worked-out examples is more effective than problem solving, which, in turn, becomes more effective at later stages (Renkl & Atkinson, 2003). This leads to the discussion about the amount and structuring of the support provided and opens the question of effective use of scaffolding to foster observational learning. Different types of scaffolding can be combined to address different sides of the observed complex skills. Although the research supports the idea of scaffolding for complex skill acquisition in general, the question of sequence and the amount of support is under question, as many different moderators should be taken into account - as, for example, students' prior knowledge or the complexity of skill to be acquired.

2.1.5 Methodological Issues in Observational Learning Research in Teacher Education: Study Design and Measurement of Outcomes.

The research on teaching utilizes both qualitative and quantitative methods and even a combination of both (hybrid studies). In line with the review by Castellan (2010), the current literature review suggests that after the middle of the 1970's qualitative research and hybrid studies have become more popular (Castellan, 2010). The literature search on Teacher Education resulted in a relatively high amount of descriptive and qualitative research with large student samples, but little quantitative research. Experimental and quasi-experimental studies in the area usually have complicated design and small samples.

Not only different presentation types, but also various measures of outcomes are used by researchers in the area. One of the most important distinction is the use of subjective (students' self-reports and self-ratings of own learning, assessment of treatment's utility and effectiveness, perceived confidence in use of new teaching techniques, etc.) vs. objective measures (teacher students' actual performance or written reflections coded and rated by experts, frequency of wanted behavior, level of skill acquisition, etc.). The distinction between subjective and objective measurement is made on methodological, theoretical and practical levels. Rothstein (1989) emphasizes that objective measures are supposed to be more reliable,

while subjective measures are more prone to errors. Although some authors consider the subjective measures to be a valuable source of information (Lee & Ertmer, 2006; Wang & Ertmer, 2003), objective measures are considered more accurate and valuable (Choppin, 1997). Moreover, Spector (1994) emphasizes the controversy of self-reported measures in measuring learning and behavioral changes.

Subjective measures reflect students' attitudes to the treatment and to their own progress; objective measures give insight on possible conceptual and behavioral changes. Due to the fact that subjective and objective measures might measure different aspects of learning it is important to analyze these measures independently from each other. Objective measures (measures of performance) in teacher education range from actual behavior to paper tests or written reflections. Actual behavior is either measured by the frequency of a specific behavior (amount of asked questions (Crooks & Gifford, 1992); reactions to students' behaviors (Kubany & Slogget, 1991; Slogget, 1972), etc.). Written tests vary from multiple choice to open-ended questions and from knowledge tests to reflections and lesson planning (Koran et al, 1971; Moreno & Valdez, 2007).

2.1.6 Bias detection and correction in a meta-analysis

Meta-analysis is a statistical analytical tool designed to summarize the results of several studies, it gives opportunity to increase the sample and, therefore, the power to investigate the effects. It helps to overcome some limitations of primary studies, and to provide higher generalizability of the results. The main methodological issue regarding any meta-analysis is that the results strongly depend on the concepts, quality and statistical power of individual studies (sample size and the design), but also to some extent on the decisions made by the researcher conducting the meta-analysis (selecting studies for the analysis, coding moderators, etc.).

The current meta-analysis on the effects of observational learning on teaching skills, , was grounded on studies with relatively small samples to combine effects reported in the

primary studies and it is essential to know if the results of these studies can be generalized and applied in the domain of Teacher Education. As only quantitative studies were utilized, the sample is relatively small. To address these issues, a range of statistical methods to control and correct for possible publication bias, questionable research practices and other manipulations was used to ensure sufficient power, validity and generalizability of the findings.

The methods used to detect and correct for possible biases are described in more detail in the method section (see section 2.3.4.2). This description includes the strengths and the weaknesses of each method and the suggestion to use the combination of these methods in order to provide a more complete and robust picture in assessing the quality and generalizability of the results.

2.2 Meta-Analysis Research Questions

To assess the effectiveness of observational learning as one of the methods in the domains of Teacher Education and the role of the presentation format, scaffolding and measures of outcomes in fostering and assessing knowledge and skill acquisition the following research questions were formulated for the current meta-analysis:

RQ1: To what extent does the observational learning in Teacher Education affect objective and subjective learning outcomes?

Observational learning is considered to have a positive effect on learning outcomes (Chi et al., 2008; Fryling et al., 2011; Stegmann et al., 2012; Van Gog & Rummel, 2010). Against the background of findings in other domains (e.g. Medical Education, Math), it seems reasonable to expect a medium to large positive effect of observational learning on subjective as well as objective learning outcomes.

RQ2: To what extent does a presentation format (in vivo, video-based or text-based) and measures of performance (actual performance or text-based skill test) influence the effectiveness of observational learning?

According to recent findings about the properties and use of video in education, video format might be more beneficial for the complex skill acquisition (Gaudin & Charlies, 2015) as it allows reviewing and focusing on the details of an authentic situation, but at the same time limits distraction. It can also be expected that the video format is effective if imitation of the behavior is required from novice teachers. In line with Bandura's theory (1986), imitation of the activity takes place during the performance phase. Therefore, if behavior is seen, it can be repeated easier than described. On the other hand, reading and discussing a case study might not automatically mean starting to use new behavior. This makes text models to be beneficial to structure experience and use the professional language, but not demonstrate the behavior and therefore be more effective if written tasks are used to measure knowledge gain. It can be assumed that both presentation format and measure of performance will be significant moderators. We further expect that measures of performance will interact with the presentation format in a way that repeating in performance phase the activity demonstrated during the presentation phase will result in a higher effect than if activities during the presentation and performance phases were different.

RQ3: To what does scaffolding influence the effect of the observational learning on learning outcomes?

As scaffolding is claimed to be essential for skill acquisition (Allen & Ryan, 1969; Darling-Hammond et al, 2005; Santagata, et al., 2007), it is expected that the more scaffolding is present for pre-service teachers, the more they can focus on the modeled behavior and, therefore, the greater knowledge gain and skill acquisition will be demonstrated. Observational learning without scaffolding is expected to be less effective as according to Stegmann et al. (2012) and Hoover et al. (2012) distractions and cognitive bias might lead to plausible explanations and misleading concepts and in turn decrease learning outcomes. We expect that scaffolding will enhance the effect of observational learning. We would also expect that if the control group in a study received no opportunity for observation, but was scaffolded in any

other way in knowledge/skill acquisition, the effect of the observational learning on the experimental group would still be significant indicating the unique value of the observation.

2.3 Method

2.3.1 Inclusion and Exclusion Criteria

2.3.1.1. Observational Learning in Teacher Education

The eligible studies were required to include at least one comparison, i.e. to compare a condition with observational learning with a condition without observational learning in teacher education. Observational learning was defined as acquiring (from observing the model) and further demonstrating of observed skill. This definition goes in line with the definition suggested by Chi (2009). Studies, which used observation as method of presenting or collecting the data, as a tool to provide feedback or assess performance, or as teaching strategy aimed at developing analytical skills, were not included in the analysis. The studies were further required to refer to teacher education. Eligible participants only included pre- or in-service teachers in different subject domains. Studies with pupils, students (from other than Teacher Education domain), parents or care-givers, employees of different organizations not connected to teaching, etc. were excluded from the current meta-analysis.

2.3.1.2. Learning Outcomes

The eligible studies were required to measure teaching-related knowledge and skills (e. g., teaching strategies or techniques (questioning, facilitating discussion, incorporating technology in the classroom), classroom management skills, attention and reactions to appropriate and inappropriate students' behavior, use of common/unique teaching methods, teachers' decision-making, reflections and observations). Studies aimed at the acquisition of content knowledge solely in any domain were not included in the analysis.

2.3.1.3. Research Design

Because the focus of the meta-analysis was to make causal inferences regarding the effect of observational learning on teaching-related learning outcomes, the studies, included in

the analysis had to have an experimental (randomized controlled trial) or a quasi-experimental design. Descriptive studies, case-studies and studies with less than four participants per condition were not included in the analysis as they violated assumptions of the statistical methods used for the analysis.

2.3.1.4. Study Site, Language and Publication Type

Eligible studies could take place in university courses, authentic classrooms (field studies), labs or other learning environments. To make sure that the concepts and definitions of the core elements coded for the meta-analysis are comparable and relevant, only studies published in English were included in the analysis. However, the studies could be conducted in different countries and not necessarily in English. Different sources were considered (project and technical reports, journal articles, conference papers, dissertations); both published and unpublished studies were included in the analysis to ensure the validity and generalizability of the results. There was no limitation on the publication year.

2.3.1.5. Effect Sizes

The eligible studies were required to report sufficient data to compute effect sizes (i. e., sample sizes, descriptive statistics) and to identify the direction of scoring. In case of insufficient data, studies required to report at least statistical values that allow an estimation of effect sizes (e. g., F, t, p, df, etc.). In case of insufficient data, the authors were contacted to get the required data. If contacting the authors was not successful and there was no opportunity to get the statistical data needed, the study was excluded from the analysis. Post-test effect sizes were used to calculate the summary effect. If a study reported the information about pre-test effect size, it was used to correct for pre-test differences between experimental and control conditions.

2.3.2 Search Strategies

The search term used to find the potentially eligible studies was “(observational learning OR vicarious learning) AND teacher”, which should be mentioned in a title or an

abstract of a study. As we focused on Teacher Education, we searched the following hosts/databases: ERIC, PsycINFO, ProQuest. The reference lists of the articles selected after the initial coding (see below) served as starting point for a "snowball" search technique. The search resulted in total k = 475 articles: 470 from databases and 5 additional articles from the "snowball" search.

2.3.3 Coding procedures

The data required for the meta-analysis was extracted from the selected primary studies. This data included study characteristics, independent and dependent variables and statistical values needed for the estimation of the effect sizes. Features of primary studies that required a higher level of inferences (study design, use of instructional support and scaffolding for control and experimental groups, presentation format and measures of outcomes) were double coded by a second coder. All disagreements between coders were solved through discussion before the analysis. Regarding statistical data, sample sizes and descriptive statistics (Means and Standard Deviations) for both pre- and post-tests, experimental and control groups, were extracted in the first step. If the study did not provide descriptive statistics, but reported correlation, regression, ANOVA coefficients, the results of the t-tests, the reported coefficients were used to estimate the effect sizes. If authors of primary studies did not report the required statistical values they were contacted via email (if possible) to get the missing values. If there was no further possibility to get the data required, the studies that provided insufficient information for calculation of effect-sizes were excluded from the analysis.

2.3.3.1. Study Characteristics

The information about authors, year and type of publication (report, dissertation or journal article), as well as the information about sample sizes (overall amount of the participants) was extracted for every study. The overview of the study characteristics and moderators are presented in Table 2.1. Some other study characteristics were used to calculate and adjust the effect sizes (type of study design, adjustment for pretest differences, correlation

between pre and post-measures, etc.). The detailed Excel table with all study characteristics can be found in the digital Appendix.

2.3.3.2. Coding of the Moderators

The first moderator introduced at the early stage of the analysis was the "Outcome Measure" (objective vs. subjective measures of learning outcome). As the subjective (self-reported utility, reported confidence in applying acquired skill or knowledge) and objective (demonstrating teaching behavior, written reflections or tests) measures are different in their nature and should not be combined, the author analyzed the studies separately.

The second moderator related to the outcome measures was "Measures of performance" (coded only for objective measures). It was coded as either (a) written measures (knowledge or skill tests, reflections); or (b) performance (an actual use of principles/behavior instances targeted as observed and rated by experts in primary studies).

The third moderator, "Presentation Format", referred to the way the model was presented to the learners. The participants could be (a) present in the classroom during real or simulated lesson (in vivo); (b) watch the recorded lesson (video); (c) read the transcript of the lesson (text) or (d) do several of above mentioned options (combined).

The fourth moderator – "Scaffolding" was coded for both treatment and control groups. For experimental groups it was coded by two separate codes: (a) during observation and (b) after observation. Scaffolding after observation (additional instructional support) referred to presenting questions for discussion, cognitive prompts for information retrieval, or similar after the observation took place to help pre-service teachers reflect on the modeled skills. Scaffolding during observation (actual scaffolding) referred to promoting active observation as defined by Chi (2009): pre-service teachers should have taken notes, rated observed behavior according to given criteria or perform similar activities during the observation. If both types of scaffolding for experimental group were used, such scaffolding

was referred to as “continuous”. “Mixed” scaffolding refers to the studies with multiple conditions and different type of scaffolding used for each of them.

Additional scaffolding for a control group was coded as (a) present (if a control group was instructed differently from the experimental group; received different treatment: lecture, pre-discussion) or (b) absent (no additional instructional support for a control group).

2.3.4 Statistical Methods

For our analyses, we followed the procedure described by Borenstein, Hedges, Higgins and Rothstein (2009) for effect sizes calculation, integration and moderator analysis.

2.3.4.1 Calculation of the Effect Sizes and Synthesis of the Analysis.

To calculate an effect size (Hedges *g*) the following steps were taken. First, the information about the study and the available statistical data were extracted from the study (sample size, descriptive statistics) for pre- and post-tests in experimental and control groups, (independently for each comparison), and inserted into the calculation sheet (Stegmann, 2015) based on formulae and recommendation suggested by Borenstein and colleagues (2009). Each comparison was coded as “pre” if it referred to the pre-test values and as “post”, if it referred to the post-test values.

Second, the study design was coded as either “independent” if the mean/scores for the two groups at the same point in time were compared (an experimental vs. a control group) or “paired” if the same group of people was tested several times (during a pre- and a post-test). For the paired design, if the authors reported the standard deviation of the difference, the study was coded as “paired_A”; if authors only reported pre-test and post-test standard deviations, the study was coded as “paired_B”. The study design codes identified which formula should be used to calculate the standard deviation of mean differences. In case of “paired” design, calculation required the correlation coefficient between pre- and post-test. If the study provided the coefficient, it was inserted directly from the study. In case the authors did not provide this coefficient, it was estimated from the studies using the same measurement scale, systematic

reviews on measures used in the research, or if no other information could be used, it was set to 0.5.

The third step (if descriptive statistics was available) implied calculating the difference between the two means (by subtracting a mean score of a control group from a mean score of an experimental group for independent design, or by subtracting a pre-test mean score from a post-test mean score for the paired design), its variance and standard error. Standard deviation of this difference was calculated using the standard deviations and the sample sizes of the groups/scores.

In the fourth step the estimated mean difference and its standard deviation were used to calculate a Cohen´s *d* coefficient (effect size). In the event that descriptive statistics was not available, Cohen´s *d* was estimated from other statistical values (for this study using *t* and *F* statistics) using the formulae suggested by Borenstein and colleagues (2009). As a fifth step, for all the studies with known pre-test differences, the effects were corrected for these differences. It was done by subtracting the pre-test effect from the post-test effect. The corresponding variances were multiplied. In step six, the correction coefficient *J* was used to adjust Cohen´s *d* for sample sizes, resulting in obtaining Hedge´s *g* for every comparison in all the primary studies.

To estimate a summary effect a random-effects model was used. Most of the studies in this meta-analysis have a complicated design resulting in multiple comparisons, but use relatively small sample sizes. One of the dangers for the current meta-analysis was the possibility of correlated samples. There are two solutions to address this problem – (1) integrate the effect size within a single study (to assure that only independent comparisons affect the summary effect calculation and moderator analysis) or (2) use the Robust Variance Estimate (RVE) suggested by Tanner-Smith, Tipton and Polanin (2016), which allows to include all comparisons and correct for possible dependencies. Both methods were tested and as RVE and the integration of effects within correlated samples produced very similar results,

it was decided to proceed with the integration of effects from correlated samples. The main reason behind this choice was the fact that moderator analyses were planned and RVE models are neither intended to provide precise variance parameter estimates, nor to test the null hypotheses regarding heterogeneity (Tanner-Smith, Tipton, & Polanin, 2016). To get a representative result only one effect size pro study was used in the synthesis. In case the study reported multiple effects, a small-scaled meta-analysis was run to synthesize the results within a single study before including the effect to the summary effect estimation.

Confidence intervals were used to assess the significance of an effect. Heterogeneity estimates (*Q*-statistics) were used to determine the variance of the true effect sizes between studies (*tau*) and the proportion of this variance that can be explained by random factors (I^2). The thresholds suggested by Higgins, Thompson, Deeks, and Altman (2003) were used to interpret the I^2 (25% for low heterogeneity; 50% for medium; 75% for high heterogeneity).

2.3.4.2 Assessment of Publication Bias

There are different ways to assess publication bias used in the research. One of the early attempts to address the sampling bias in the literature search was the fail-safe N method suggested by Rosenthal (1979). The fail-safe N reflects the number of additional studies with zero effect that would be needed to increase the p-value of a meta-analysis study to above 0.05. In other words, it assesses the number of studies that would need to exist to make an overall effect size non-significant, either statistically (Rosenthal, 1979) or practically (Orwin, 1983). This method can be used to estimate the significance of the results obtained from the sample of studies and their generalizability. One of the main limitation of the method is that it underestimates the complexity of a meta-analysis and assigns all possible variance to sampling error (Orwin, 1983). Therefore, on the one hand, the method provides important information, but, on the other hand, needs to be combined with other methods to consider the statistical model underlying the meta-analysis.

Another set of methods (Egger´s test, Trim´n´fill) is based on a graphical representation (scatter plot) of the relationship between the effect sizes and their standard error. Egger's test in the absence of publication bias, assumes that studies will be spread evenly on both sides of the average, but if publication bias is present, reported effect sizes correlate with sample sizes (Sterne & Egger, 2001). Therefore, an uneven distribution (with most studies on the right side from the average) would be a signal of a publication bias. The outliers and studies clustered together can also indicate an additional source of variance that needs to be considered. To assess symmetry of funnel plot Trim'n'fill techniques is usually used, as it detects the publication bias, and also suggests the correction. If publication bias (most values on the right side of the funnel) is detected, missing values are added on the left side of the plot to correct for bias (Duval & Tweedie, 2000) and estimate the true effect. But this estimation would only be valid if a publication bias is only due to the effect size, not by statistical significance (Duval & Tweedie, 2000). Another important limitation of the funnel plot based methods is that they disregard that the true effect sizes might differ across studies, and therefore might conclude publication bias even if it is not present (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006)

The p-curve analysis is one of the more recent methods that addresses both detection and correction for possible publication bias and evaluates the significance of estimated effect sizes (Simonsohn, Nelson, & Simmons, 2014). It also enables detection of possible questionable research practices, which are difficult to capture with the fail-safe N technique. This technique provides a robust estimate of the significance of p-values from the studies, plots them and combines the half and full p-curve to make inferences about an evidential value (Simonsohn, Simmons & Nelson, 2015). Evidential value means being authentic and relevant evidence of true effect, unbiased and free from questionable research practices e.g. p-hacking (uncovering patterns in data that can be presented as statistically significant).

L-shape distribution with most of the values below .025 is an evidence of evidential value. Flat line (even distribution of p-values from 0.01 to 0.05 is an evidence of no effect, inverted L-shape distribution (most of the values are around 0.05 with little or no values at 0.01 is an evidence of publication bias and/or questionable research practices used. Figure 2.1 was created using the online tool created by Simonsohn and colleagues (2014) to demonstrate possible distributions of the p-values. The x axis represents the p-values (from 0.01 to 0.05) and the y axis the percentage of studies reporting these p-values.

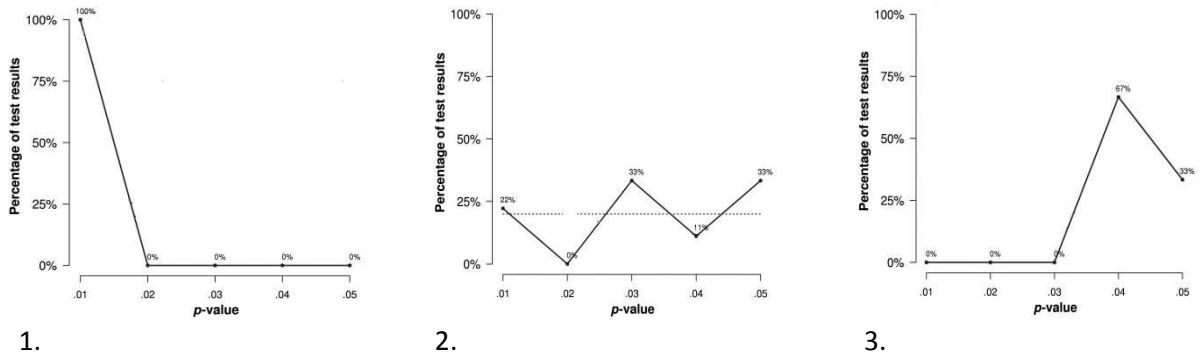


Figure 2.1. Distribution of p-values: 1. Evidential value, 2. No effect, 3. Publication bias

To enable p-curve analysis, p-values should test the hypothesis of interest (only p-values relevant to the hypothesis are estimated), have uniform distribution and be independent from other p-values, therefore only one p-value was estimated per study. The method also has several limitations. First, implementing of p - curve statistics allows estimating and correcting for publication bias using only significant studies (below 0.05), but excludes studies close to being significant. Second, it also might fail to provide adequate evidential value in quasi-experimental studies, or when a covariate correlates with the independent variable of interest (Simonsohn et al, 2014).

One more method that takes under consideration both significant and insignificant results is R-index. It can be used to examine the credibility and replicability of studies, in other words, to predict whether a set of published results will replicate in a set of exact replication studies (Schimmack, 2012). According to Schimmack (2016) the R-index can be between 0

and 100%; values below 22% indicate the absence of true effect, and values below 50% indicate inadequate statistical power of the study; values above 50% are acceptable to support credibility and replicability of the results, although values above 80% are preferred. There are also some issues to be considered about the method. The R-Index builds on the incredibility index and the probability provides no information about effect sizes and amount of studies in the analysis (Schimmack, 2016).

The strategies used to detect and estimate publication bias have different assumptions and limitations. As current research relies on heterogeneous studies with complex design and relatively low sample size, it takes under consideration multiple techniques as their combination allows assessing and quantifying publication bias and questionable research practices in the most precise way. If the results of different methods contradict each other, it would raise further questions about which method is the most adequate under given conditions, but if the results of all methods applied to test for publication bias go in the same direction, it would be a strong indicator either for or against publication bias.

2.4 Results

2.4.1 Results of the Literature Search

In the first step, the abstracts and titles of these 475 articles were read and in accordance with inclusion criteria, 123 articles were selected for further coding. The most common reasons to exclude the studies from the analysis were either due to the qualitative / descriptive nature of the study or fostering the acquisition of skills not connected to teaching. In the second step, the articles selected from coding abstracts were thoroughly read and coded in terms of methods and variables used. The most common reason to exclude the article from further analysis was mentioning/using observation techniques for data collection or feedback, rather than for skill and knowledge acquisition. Fifteen articles reporting 19 independent studies satisfied the inclusion criteria and reported statistics needed for estimation of effect sizes. From the 19 independent studies, 13 reported effects based on objective measures (total

of 84 mean comparisons included, of which up to 16 independent comparisons were included in subsequent moderator analysis); and 6 studies reported effects based on subjective measures of learning outcomes (11 mean comparisons included). The selection process and the results of literature search are presented in Figure 2.2.

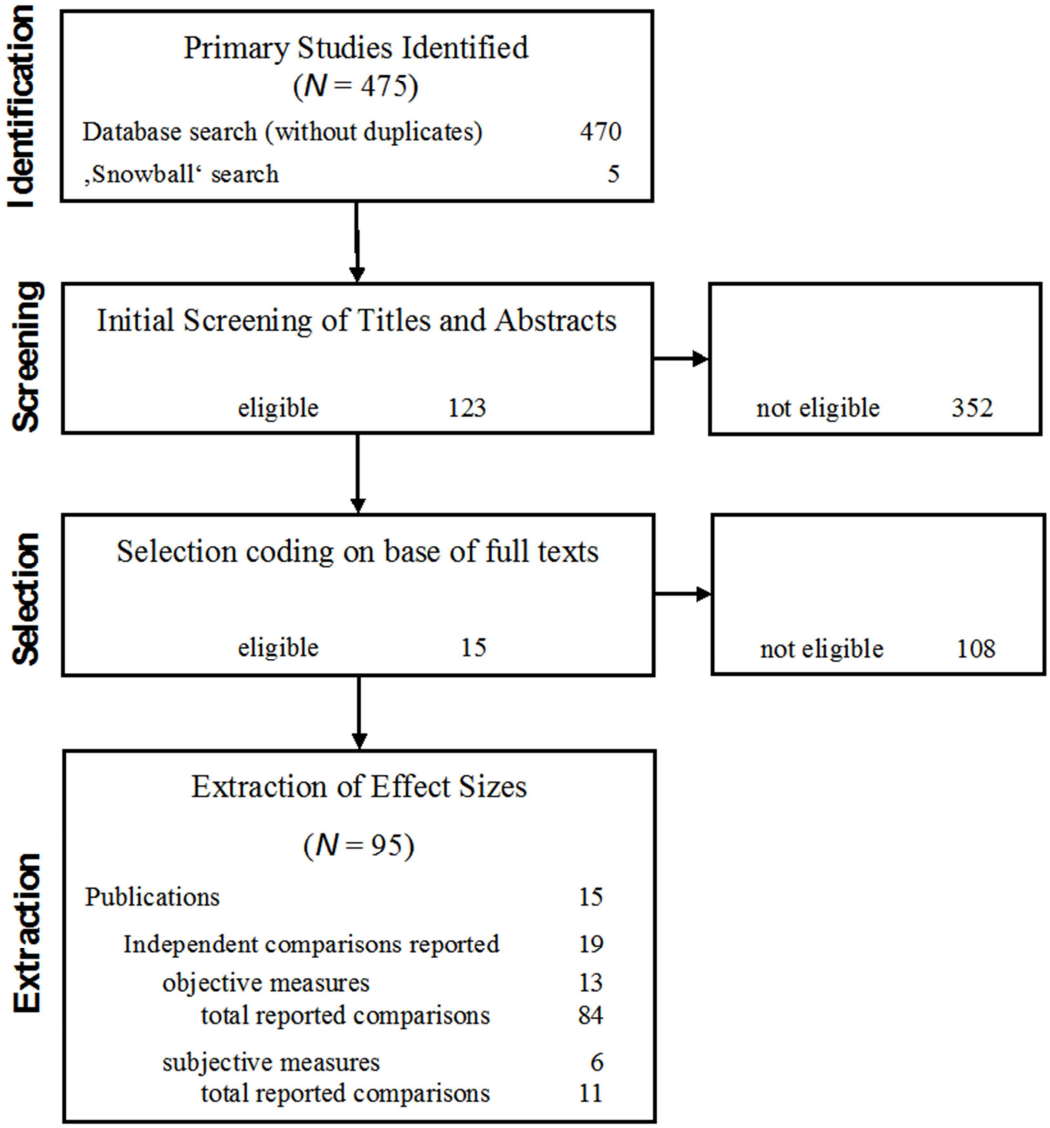


Figure 2.2. Study identification and effect size extraction flow diagram.

The final sample of studies used in the analysis is presented in Table 2.1. The table includes information about publication year and type, amount of participants and the coding of moderators.

Table 2.1

Summary information about studies in the meta-analysis

Study	Paper	Publication type	N	Measures of outcome	Posttest	Present. format	Scaffolding
1.	Bloch, 1977	dissertation	46	Objective	teaching	video	no scaffolding
2.	Claus, 1969	tech. report	40	Objective	teaching	video	continuous
3.	Crooks & Gifford, 1992	j. article	38	Objective	teaching	video	after observation
4.	Gettinnger & Stoiber, 2014	j. article	6	Objective	teaching	in vivo	no scaffolding
5.	Haverback & Parault, 2011	j. article	86	Subjective	rating	video	after observation
6.	Koran Jr. et al., 1969	j. article	33	Objective	written	video	mixed
7.	Koran Jr.et al., 1970	j. article	21	Objective	written	video	no scaffolding
8.	Koran et al., 1971	j. article	120	Objective	written	video/text	after observation
9.	Koran Jr. et al., 1972	j. article	78	Objective	teaching	video	no scaffolding
10.	Kubany & Slogget, 1991	j. article	61	Objective	teaching	video	mixed
11.	Lavin, 1992	dissertation	48	Objective	teaching	in vivo	continuous
12.	Lee & Ertmer, 2006	j. article	65	Subjective	rating	mixed	after observation
13.	Moreno & Valdez, 2007; exp. I	j. article	53	Objective	written	video/text	no scaffolding
14.	Moreno & Valdez, 2007; exp. I	j. article	53	Objective	written	video/text	no scaffolding
15.	Moreno & Valdez, 2007; exp. II	j. article	55	Subjective	rating	video/text	no scaffolding
16.	Moreno & Valdez, 2007; exp. II	j. article	55	Subjective	rating	video/text	no scaffolding
17.	Slogget, 1972	dissertation	67	Objective	teaching	video	mixed
18.	Wang & Ertmer, 2003; exp. I	conf. paper	20	Subjective	rating	mixed	no scaffolding
19.	Wang & Ertmer, 2003; exp. II	conf. paper	20	Subjective	rating	mixed	after observation

Note: *Multiple experiments within one paper were treated as independent studies.*

2.4.2 Summary effects of observational learning (RQ1)

2.4.2.1 Summary effect on objective measures of learning

The integrated adjusted effect size from 13 studies reporting comparisons regarding objective measures of observational learning on teachers´ learning outcomes is g = 1.13, SE = 0.21, 95% CI [0.72, 1.54], $p < .001$, indicating statistically significant large positive effect of observational learning on learning outcomes. The analysis also showed high heterogeneity, $Q(12) = 149.48$, $p < 0.001$, $\tau^2 = 0.48$; $I^2 = 91.97\%$, therefore further moderator analysis was performed. Figure 2.3 displays the individual effect sizes for the studies in forest plot.

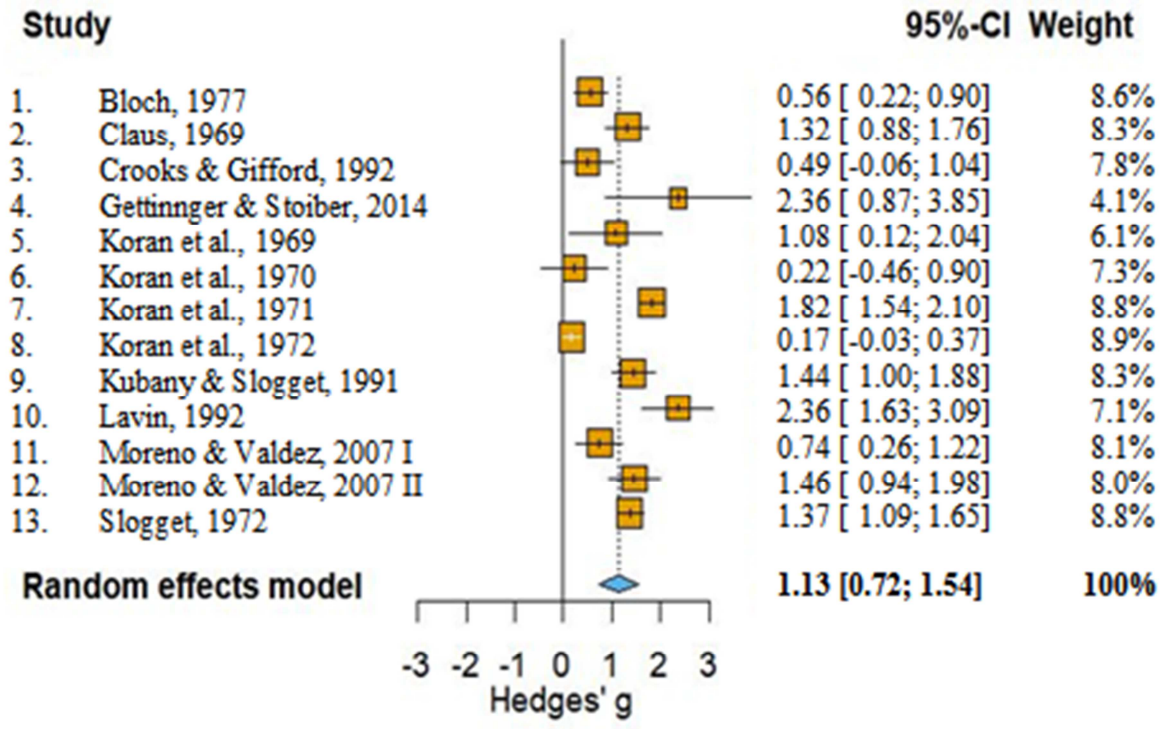


Figure 2.3. Forest plot for the observational learning on objective measures of learning.

2.4.2.2 Summary effect on subjective measures of learning

The summary effect from 6 studies reporting comparisons regarding subjective measures of teachers´ learning is g = 1.07, SE=0.24, 95% CI [0.60, 1.54], $p < .001$, indicating statistically significant large positive effect of observational learning on learning outcomes. The analysis also showed low to medium heterogeneity of effect sizes, Q (5) = 8.60, $p < 0.001$,

τ^2 = 0.13, I^2 = 41.87%. Figure 2.4 displays the individual effect sizes for the studies in forest plot. The relatively homogeneous effect sizes in combination with the small sample size led to the decision to exclude effects of observational learning on subjective measures from the further moderator analysis.

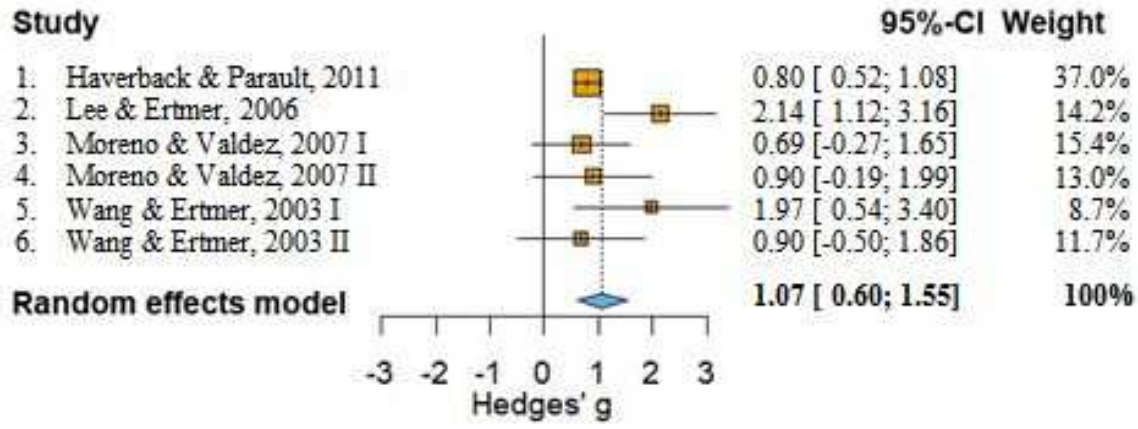


Figure 4. Forest plot for the observational learning on subjective measures of learning.

2.4.2.3 Estimation of publication bias for summary effects

The funnel plots for effects on teachers´ learning measured by objective and subjective measures in Figure 2.5 provide an indication that the sample of studies selected for meta-analysis is not affected by publication bias. For both objective and subjective measures Trim'n'fill technique also supports the absence of publication bias, no studies are missing on the left side of the funnel plot. Fail-safe N suggests that the number of additional studies with zero effect that would be needed to increase the p value for the meta-analysis to above 0.05 is 993 studies for objective measures and 78 for subjective, which goes in line with the size of the effect and power estimation of the meta-analysis.

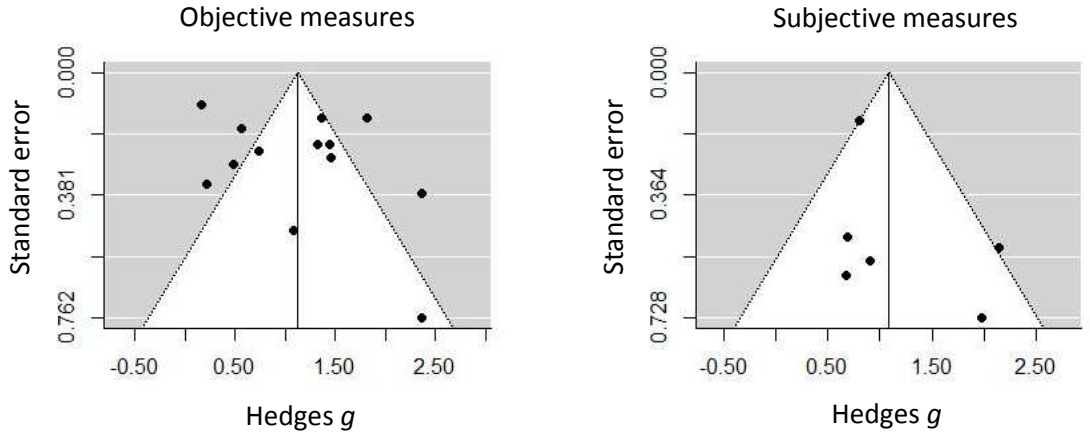


Figure 2.5. Funnel plots: the summary effects of observational learning on teachers´ learning.

Figure 2.6 presents the results of implementing p-curve analysis, which support the claim that the studies contain evidential value (according to continuous tests for the whole and the half of p-curve, $p < 0.01$). For objective measures 11 statistically significant ($p < 0.05$) results were used to create the curve, 9 of them were below $p < 0.025$. Two results were not used as they were nonsignificant ($p > 0.05$). For subjective measures the observed p-curve includes three statistically significant ($p < 0.05$) results, all of them below $p < 0.025$. Three nonsignificant results were not used to create the scale.

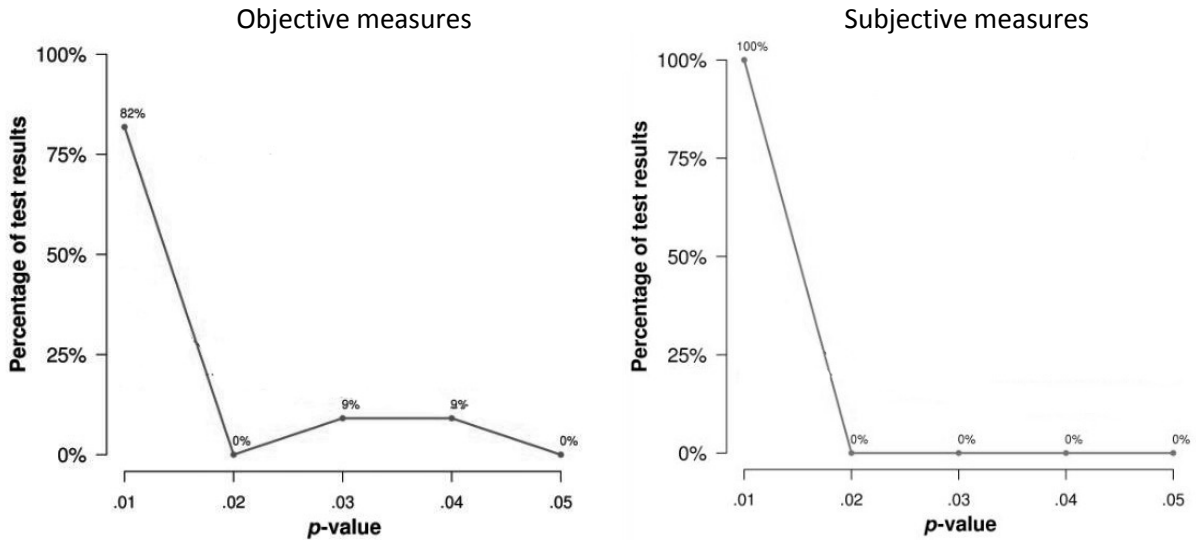


Figure 2.6. P-curves for the summary effects of observational learning on teachers´ learning.

The R-index was calculated for all the studies to assess the replicability and power of the effect sizes used for the analysis. It assessed the probability to find the reported effect if a study is replicated, in meta-analysis it assesses the probability of replication studies to find the reported effect. The R-Index for objective measures is presented in Table 2.2. Six studies had 80% and higher R-index and were significant. From five studies with R-index below 50%, two studies were non-significant. The R-index for subjective measures is presented in Table 2.3.

Table 2.2

Replication indices for objective measures of teachers' learning.

Study	Observed power	p<.05 (%)	Inflation Rate	R-Index (%)
Bloch, 1977	0.48	50	0.02	47.00
Claus, 1969	1.00	100	0.00	100.00
Crooks & Gifford, 1992	0.48	50	0.02	46.00
Gettinnger & Stoiber, 2014	1.00	100	0.00	100.00
Koran et al., 1969	0.61	100	0.38	22.00
Koran et al., 1970	0.08	0	0.08	15.00*
Koran et al., 1971	1.00	100	0.00	100.00
Koran Jr. et al., 1972	0.16	0	0.16	32.00*
Kubany & Slogget, 1991	0.81	83	0.02	79.00
Lavin, 1992	1.00	100	0.00	100.00
Moreno & Valdez, 2007 I	0.52	50	0.02	54.00
Moreno & Valdez, 2007 II	0.92	50	0.42	100.00
Slogget, 1972	0.82	83	0.02	80.00
Average	**0.68**	**67**	**0.09**	**67.31**

Note: * reported non-significant results.

Table 2.3

Replication indices for subjective measures of teachers' learning.

Study	Observed power	p<.05 (%)	Inflation Rate	R-Index (%)
Haverback & Parault, 2011	0.75	75	0.00	75.00
Lee & Ertmer, 2006	1.00	100	0.00	100.00
Moreno & Valdez, 2007 I	0.98	100	0.02	97.00
Moreno & Valdez, 2007 II	0.21	0	0.00	41.00*
Wang & Ertmer, 2003 I	0.87	100	0.03	94.00
Wang & Ertmer, 2003 II	0.25	0	0.00	50.00*
Average	**0.68**	**62.5**	**0.00**	**72.83**

Note: * reported non-significant results.

2.4.3 The Role of the Presentation Format and the Measures of Performance (RQ2)

The presentation format, the measures of performance and their interaction were analyzed to address the heterogeneity of the effects of observational learning. The results of the Q-tests for both moderator variables support the assumption that the presentation format, $Q(2) = 14.26$; $p < .01$, and the measures of performance, $Q(1) = 6.32$; $p < .01$, were statistically significant moderators for the effect of observational learning on learning outcomes measured with objective measures.

Presenting (observing) models in vivo resulted in a significant positive effect of observational learning on the objective measures of learning outcomes. There were only two studies which reported using this presentation format. This provides insufficient evidence about the effectiveness of this type of observation and does not allow the assessment of the combination of presenting and measuring formats. Learning with text and video models showed comparable effects (see Table 2.4).

Pre-service teachers utilized the observational learning similarly well for the actual performance and written measures. Measuring actual performance at the post test appeared to indicate more gain than if written measures were used, but the data was insufficient to make conclusions about the best way of measuring the knowledge gain. The combination of text presentation and written measures yielded similar results to the combination of video presentation with actual performance (see Table 2.4).

Although there was some heterogeneity left within the groups, the included studies were not enough and provided insufficient data to identify other possible moderators. Considering the identification and correction for publication bias and the questionable research practices, no biases were detected. The continuous test for p-curve analysis supported the evidential value of the results and the replicability indices were acceptable for all of the moderator levels.

Table 2.4.

Results from the Moderator Analyses Examining Differences in the Adjusted Post Test Mean Effect Sizes for Observational Learning to Non-Observational Learning Conditions.

Moderator variable	*p*	*g*	95 % *CI*	*n (k)*	*EV*[1]	*R - index*	τ^2	I^2
Presentation format	*							
Video		1.06	[0.50, 1.62]	494 (11)	1	63%	0.82	95.98%
Text		1.09	[0.59, 1.59]	152 (3)	1	99%	0.13	68.36%
Presence		2.36	[1.71, 3.02]	30 (2)	1	-	-	-
Measures of performance	*							
Performance		1.24	[0.73, 1.74]	408 (9)	1	65%	0.52	93.25%
Written		1.10	[0.50, 1.69]	194 (5)	1	82%	0.36	81.19%
Interaction of presentation format & measures of performance	*							
Performance + video		1.08	[0.45, 1.71]	378 (7)	1	62%	0.69	96.29%
Performance + text		-	-	78 (1)	-	-	-	-
Written + video		1.31	[0.43, 2.20]	196 (5)	1	69%	0.87	87.24%
Written + text		1.10	[0.46, 1.73]	152 (3)	1	98%	0.19	61.15%
Scaffolding	*							
Not after. & During		1.10	[-0.12, 2.31]	174 (3)	1	67%	1.09	95.01%
Not after.; not during		0.85	[0.48, 1.21]	342 (9)	1	59%	0.21	78.71%
After & During		1.80	[0.78, 2.82]	87 (2)	-	-	-	-
After; Not during		1.30	[0.39, 2.21]	141 (3)	1	88%	0.56	88.70%

Note: * $p < .05$ significant;
[1] evidential value: 1= significant ($p < .01$); 0 = nonsignificant ($p > .05$);
τ^2 = between studies variance component

2.4.4 The Role of the Scaffolding (RQ3)

Table 2.4 displays the analysis results of the effect of two types of additional instruction (scaffolding) and their interaction. The first group illustrated scaffolding only during observation (i.e. using cognitive prompts, observation protocols or similar tools to focus the students´ attention on the core elements of the observed model); the second group illustrated results for using no scaffolding at all; the third group illustrated use of continuous scaffolding (during and after observation); and the forth group – additional instructions after observation (discussing the model observed, answering the questions to the observed model, writing reflections or similar activities).

From the results it is noticeable that continuous scaffolding was the most effective type of support. It also had the highest replicability index possible, which in line with the previous research findings, provides the supportive evidence for the claim that support is need as much possible at least in the initial stages of learning for pre-service teachers. Absence of any scaffolding was the least effective. Two types of scaffolding (provided either after or during observation) were comparable when it comes to the size of the effect; however the studies included in the analysis indicate that the effect of scaffolding during the observation should be interpreted with caution as it might not be considered significant (confidence intervals include zero). The wide confidence intervals and the high heterogeneity for the studies with scaffolding during the observation might be due to the differences in the methodology of the studies. Although the heterogeneity stayed high within groups, further moderator analysis was not possible due to insufficient data. The evidential value of the results was supported by the continuous test in p-curve analysis; and the replicability index was acceptable for the three of four levels of moderator. The fourth level provided insufficient data for calculation of the R-index.

Scaffolding given to control group failed to become a significant moderator: there was insufficient variance between two levels of moderator (scaffolded/unscaffolded control group),

Q (1) = 1.57, p = 0.15, ns. Moreover, no conclusions could be drawn for the condition when the control group received any kind of instruction different from observation (lecture, pre-discussion), which was not presented to the experimental group. Although the effect of observational learning on learning outcomes remained significant, indicating the added value for pre-service teachers' acquisition of knowledge and skills, evidential value of these results is not supported with p-curve analysis (continuous test for evidential value p > 0.05).This might be an indicator of insufficient amount of studies or that the results appeared by chance.

2.5 Conclusions

The current meta-analysis was conducted on the set of 19 studies measuring the effect of observational learning on learning outcomes measured by objective measures (13 studies) and subjective measures (6 studies) to address the size and magnitude of summary effect of observational learning and some moderators affecting the knowledge gain and skill acquisition (presentation format for the model, measures of outcome, scaffolding).

Due to methodological issues a range of methods to assess possible publication bias, power and replicability of the study were applied. In the initial stages, studies using subjective vs. objective measures were analyzed separately, as they could have measured different aspects of learning. Although the integration of the effects for both subjective and objective measures yielded similar results, the effects of the studies were not combined for further analysis, to avoid integrating conceptually different outcomes.

Although 6 and 13 studies is a relatively small sample for the meta-analysis, the results indicated that as large positive effects are estimated and no publication bias or data manipulation detected, the current sample size was enough to draw conclusions about the summary effects for this meta-analysis. Results of the moderator analysis opened a call for more research in the area to clarify the role of instructional support and its amount to enhance learning acquisition.

In line with cognitive and social-cognitive theories (Bandura, 1986; Huitt, 2004; Sweller, 2005) and recent research (Chi, Roy & Hausmann, 2008; Fryling et al, 2011) that claims observational learning to be effective instructional for the acquisition of complex cognitive skills, current meta-analysis supported observational learning to have high effect on learning outcomes (both subjective and objective).

Presentation format and measures of performance proved to be statistically significant moderators, but the size of the effect was similar between different groups. This could be due to relatively small sample sizes that led to increased variance of the effects. Observation in vivo (direct observation) seemed to be a very promising approach and therefore needs further exploration. On the other hand, despite its effectiveness it might be ethically and technically difficult to use the in vivo approach to educate a big amount of pre-service teachers. The main idea of direct observation is not to alter the classroom environment so no more than 1 or 2 pre-service teachers can observe a situation in vivo if no technology support is used. Video and text models according to the current meta-analysis might do similarly well, as they allow for focusing on details, re-watching important moments and reflecting on them, in other words share strengths and can be used interchangeably. This is particularly important for planning research, as it allows to reduce the resources and still reach similar effectiveness. The findings also indicate that if presentation and measurement format are combined as moderators, the results are difficult to interpret in favor for one good combination, but it shows that observational learning can work in different circumstances.

It is worth noting, that although many researchers (Chi et al., 2008; Dianovsky, & Wink, 2011; Glogger et al., 2009; Hübner, 2009; Van Gog & Rummel, 2010) emphasize that observational learning is only effective if properly scaffolded, results of this meta-analysis show that the effect stays medium even if observational learning is not additionally scaffolded. This might be due to the fact that researchers did not report all the additional instructions given

to students in the course of the experiment. It could also mean that even a brief instruction is enough for students to start learning from the model.

On the other hand, in line with the theoretical framework and the empirical findings, using scaffolds might significantly increase the learning gain from observation by removing distractions, misinterpretations or other counterproductive phenomena. Scaffolding only during observation had mixed results and could have been even distracting or overloading for pre-service teachers. In other words, scaffolding during observation without meaningful follow-up is a lost opportunity to get the most from observational learning. When compared with other instructional methods (e.g. lecture) observational learning still might have additional value, but as there are insufficient data to reach the evidential value for this comparison, more research is needed before any conclusions can be drawn.

3 Measuring Lesson Planning Competency: The Scale Development

The third chapter presents the theoretical framework, some methodological issues, the procedure, and the results of the scale development to measure a lesson-planning competency in elementary school physical education pre-service teachers. The chapter starts with the general introduction to the theory and the research on competence and then deepens into the teaching competence in more detail. The boundary approach was implemented to define teaching competence and in particular the lesson-planning competency as its essential part. The relationship between lesson planning competency and other concepts related to teaching competence (professional vision and teacher decision-making) were also discussed to better define and specify the measurement model.

Although recent research suggests several dimensions to be considered in measuring competence, the study adopts the unidimensional approach, as it focuses on the rather specific part of the teaching competence. Theoretical framework and research findings provided in the first section allowed formulating goals and hypotheses for scale development and defining the criteria to select the statistical methods.

The method section starts with an introduction to the item response theory (IRT) approach and its advantages in designing a scale for measuring a competence in general and the lesson planning competency in particular. Methodological issues and assumptions of the IRT approach are discussed to address the needs and hypotheses of the current study. Following the justification of the selected statistical method, the sample, design, and procedure of development and testing the measurement tool are described. The process of scale development is presented together with intermediate results to justify the decisions about removing irrelevant or unfair items.

The result section includes the description of the final scale and estimation of its reliability and validity. The chapter ends with the discussion on some limitations and further

implementation of the measurement scale. The measurement scale is implemented in the empirical study (see Chapter 4).

3.1 Problem statement and the theoretical backgrounds

Over the past two decades, several studies have identified some methodological and practical issues with measuring teaching competence and skills in research (Ingvarson & Rowe, 2008; Van Der Vleuten, 1996). For example, several researchers, who focus on the measurement of competence (Epstein & Hundert, 2002; Koeppen, Hartig, Klieme, & Leutner, 2008), performed explorative and descriptive studies, that provided information on the content and the context of teaching, but did not assess the quality of teaching or suggest interventions to foster the development of the competence. The quality assessment would be beneficial for developing programs and courses to educate teachers. In contrast, several other studies introduced interventions to foster specific teacher skills (e.g. a classroom management, presentation, setting up working environment, etc.), most often based on a single didactic principle (Crooks & Gifford, 1992; Koran Jr., 1969; Koran Jr., 1970; Slogget, 1972) such as teaching a very specific topic or using a specific teaching technique.

As a teaching competence is not merely an addition of separate skills, measures of acquisition of single didactic principle can hardly provide enough information to measure and assess the level of teaching competence or provide insights for development and improvement. Therefore, the current chapter is dedicated to the development of a measurement scale that can close the gap between assessing local effects in terms of the acquisition of very specific teaching skills and assessing a level of teaching competence as a more general measure of knowledge and skill acquisition. The boundary approach to competence (Stoof, Martens, Van Merriënboer & Bastiaens, 2002) was used to define the model of teaching competence and develop the measure, The Rasch scale procedures were used to establish a general assessment of teaching skills contributing to the teaching competence.

3.1.1 The Boundary Approach in the Definition of Competence

Competence is a widely used concept in research assessing the quality of education and to identifying challenges. Shavelson (2010) defines the competence as a combination of physical or intellectual skills and/or abilities, conceptual and procedural knowledge to enable performance in a standardized situation, which is at the same time authentic, assessed by some level of the standard. Competence can be learned or improved and it draws upon an underlying complex ability. According to the most recent review of Blömeke and colleagues (2015) any definition of competence should involve complex cognitive skills and abilities along with affective and volitional factors, which allow the competence to work in situations of interest. Blömeke and colleagues (2015) suggest approaching competence as a continuum, a process of realizing knowledge, skills, and motivation into performance rather than merely dichotomous construct. The present study will focus on the development of the cognitive part of teaching competence, taking into consideration that motivational factors should be assessed independently during developing teacher programs and empirical studies.

The term “competence” is often used in psychology, education and other disciplines. There is some common understanding of teaching competence. Nevertheless, definitions and approaches vary significantly between researchers, that is, competence can be interpreted either as standardized requirement, intended outcome of learning, measure of ability or as a process leading to performance or even the performance itself (see Table 3.1).

Table 3.1

Definitions of competence

Authors	Definitions
Carroll, 1993	Competence as general dispositional construct, intellectual abilities independent from context
Chomsky, 1968; Patel et al., 1996;	Competence as a general set of complex cognitive skills that can be modeled and learned independently from specific situations.
McClelland, 1973; Weber & Westmeyer, 1998; Weinert, 1999; Koeppen et al., 2008	Competence as a context specific construct, defined by the range of situations and tasks to be mastered, interaction between situational challenges and the abilities of the person.
White, 1959; Epstein, 1973;	Competence as a combination of knowledge,cognitive skills and motivational factors, competence as part of self-concept.
Blömeke et al., 2015	Competence as a continuous process connecting dispositional and motivational factors, specific knowledge and skills with the performance in specific situations

Therefore, the first step to develop a measure is to define a specific competence. This can be done in several ways, but one of the most promising approaches is the boundary approach to competence proposed by Stoof, Martens, Van Merriënboer and Bastiaens (2002). Stoof and her colleagues emphasize that it is hardly/barely possible and not even necessary to come to a single true definition of competence. Instead they proposea constructivist approach to come to the variety of competence definitions fitting the context in which each of them is used. To do so, they suggest two techniques: inside-out and outside-in.

The *inside-out* technique assumes that the definition of competence should be formulated by identifying its position along several dimensions (not necessarily only on the extremes): with a focus on personal vs. task characteristics, individual vs. distributed competence, specific vs. general competence, competence as a level or competence having different levels within it, and teachable vs. not teachable competence. Identifying the position of the competence definition along these dimensions is an important step to define the

competence, its elements, and application for research or practice. Moreover, apart from identifying the competence structure from the inside, it is very important to emphasize the difference between the competence and related terms. This is the *outside-in technique* and contrasts the competence with terms such as performance, qualification, ability, expertise, knowledge, skills and attitudes, etc.

The second step would be to formulate the expectations for each level of competence, namely what would make a better performance. For example, would the pre-service teachers with higher competence level notice more details during the observation use specific depth of elaboration or use more terminology, compared to pre-service teachers with a lower level of competence. The third and final step would be to design items to measure the competence and pilot them on pre-service teachers.

3.1.1.1 Competence vs. knowledge and complex cognitive skills

Competence and cognitive skills are strongly related to each other. Cognitive skills (together with knowledge and motivational factors) are considered the building blocks for competence. Stoof et al. (2002) emphasize that competence is more than the sum of domain specific knowledge, skills and motivational factors; these building blocks are strongly interconnected but have different weights in different situations and also meta-cognitive knowledge and skills come into play.

Motivation, attitudes and beliefs play a large role in how knowledge and skills are obtained, connected and lead to performance. Having more knowledge or higher skills (i.e., in self-regulation) might influence motivation in a way that teachers spend more effort or are more satisfied with their work, or instead might rather choose the strategies they feel confident with. However, in general, motivational factors such as values, attitudes, and beliefs are difficult to change (Lai, 2011), and might thus become a barrier to develop skills and obtain knowledge, instead of serving as a good predictor of teacher performance. It is important to address the teaching beliefs and motivation in assessment, but firstly, it can be problematic to

obtain/construct acceptably objective and reliable measures. Secondly, the rather stable nature of attitudes and beliefs make it difficult to change them, the focus should rather be on teachable and changeable components of competence.

Knowledge is another building block of a competence. Teacher professional knowledge consists of (1) general pedagogical knowledge, (2) subject matter (or content) knowledge, (3) pedagogical content knowledge, and (4) knowledge of context (Grossmann, 1990). The concept of pedagogical content knowledge is close to teaching skills and teaching competence, as it includes both knowledge of content ("what") and knowledge of pedagogical principles to deliver this content ("how"). Nevertheless, knowing "what" and "how" does not automatically lead to effective performance and therefore should not be used interchangeably with a concept of competence.

Knowledge is important as it defines content to be taught, goals and objectives as well as teaching strategies and methodology to be used. Declarative knowledge is relatively easy to measure in a reliable and objective way, for example, in a multiple-choice test, but knowledge itself does not lead to performance. The meta-analysis by D'Agostino and Powers (2009), reports that test scores are low to moderately related to teaching competence and the performance in preparatory programs (application of knowledge and skills obtained in training) was a better predictor of performance. Knowledge should be applied to a situation, and it is where cognitive and meta-cognitive skills come into play. A high level of cognitive skill is associated with related knowledge and indicates that it is integrated and can be flexibly used in different teaching situations.

Therefore, complex cognitive skills are the building block with more weight, because this can be diagnostic to assess competence in general and is directly connected to performing the activity and therefore more observable. Due to their interrelation with other components, cognitive skills might identify problems with other components and itself. For example, if there is a lack of knowledge, motivation, or cognitive skill, the assessment can show this, provided

that the measurement scale for the assessment considers all factors of interest and measure them in an objective and reliable way.

Although this dissertation focuses on cognitive skills to make inferences about teaching competence, it does not mean that competence is perceived as narrowed down to the single skill. The simplification is only for the purpose of measurement and it can be partly overcome if not a single cognitive skill, but a more complex construct (including underlying knowledge and controlling for attitudes and beliefs) would be measured and analysed. See Chapter 3 for more detail.

3.1.1.2 Competence vs. competency

The terms “competence” and “competency” both appear in research and are sometimes used interchangeably. Blömeke et al. (2015) emphasize that although competency can be part of competence, both terms have a similar structure (i.e., a combination of knowledge, skills and attitudes).The term ‘‘Competence’’ (plural ‘‘competences’’) is the broader term of the two and used in holistic approaches, whereas the term ‘‘competency’’ (plural ‘‘competencies’’) is used in analytic approaches and is considered to be a part of competence. Another point of view is that competency is used more in relation to task characteristics (what are the elements of the tasks to be performed to perform effectively?), but competence is more related to personal characteristics that lead to superior performance (Stoof et al., 2002). Regardless of the differences, both competence and competency are regarded as learnable and can thus be improved (Epstein & Hundert, 2002; Shavelson, 2010; Weinert, 2001).

To conclude, researchers treat competence as the more general construct, and competency as the narrower construct focusing more on specific task. This dissertation uses the approaches of Stoof et al. (2002) and Blömeke et al. (2015) to develop a scale to measure lesson planning competency as a more specific construct to enable future empirical studies which in turn can allow to make inferences about effective instructional methods that can be used to foster teaching competence as more general construct.

3.1.2 The teaching competence: definition and core elements

Hunter (1976) defines teacher competence as the combination of "what" (intended objective) and "how" (using appropriate principles and techniques). In their systematic literature review Gaudin and Chalies (2015) also emphasize that in teacher education the knowledge and skills that are typically trained are selective attention, knowledge-based reasoning, building knowledge on "how to interpret and reflect", and building knowledge on "what to do".

As mentioned in the General Introduction, this dissertation uses the term "teaching competence" to emphasize the focus is on activities performed by a teacher, rather than on stable teacher characteristics. Using Stoof et al. (2002) definitions of competence dimensions, the focus of teaching competence is rather on specific, distributed teachable competence that can have different levels (low or high) and on task characteristics (tasks to be performed for teaching to be effective). The research in Teacher Education addresses the whole range of theoretical and practical questions and uses a number of related terms: teacher qualification, performance, expertise, teaching (or teacher) competence, etc.

Teacher qualification is associated with obtaining a teaching degree according to national standards (e.g., certificate, diploma or similar). It represents proof that a person has the knowledge and the skills to teach. Competence however can exist even prior to a formal qualification.

Performance can be defined as observable behavior. Performance can be effective, average, or poor, although poor performance is mostly never associated with qualification or competence. To contrast, competence is usually not observable by itself, but rather an underlying prerequisite for efficient performance. Successful performance in teaching is that teachers set learning goals in such a way that they addresses learners' needs (e.g., goals to foster skill acquisition, provide safety, challenge and positive attitudes towards discipline).

Expertise, according to Herling (2000), refers to optimal efficiency (given that competence is minimal efficiency). Expertise is what distinguishes between experts and novices – highly efficient performance as a result of applied skills, knowledge and experience rather than chance.

Teaching competence is a product of (1) pedagogic and domain specific conceptual and procedural knowledge, (2) complex cognitive skills and (3) motivational and affective factors that lead to effective planning and successful performance in classroom situations. As a product of these three elements, teaching competence can only lead to effective performance if all elements have a positive value. In other words, motivational factors play an important role for successful performance, as without motivation (zero or negative value) knowledge and skills will not be applied, without knowledge or skills, motivation will not lead to success, without specific knowledge the content of the lesson will be missing.

3.1.3 Lesson planning competency

Lesson planning is an essential part of teaching competence as it provides a structuring and organisational aid to the teacher. Through planning teachers decide about specific skills and competencies that need to be fostered or acquired by the students, teaching and learning strategies that can be used, and measures to assess the effectiveness of the acquisition of the skills and competencies targeted. Lesson planning can take different forms: it can be written for each separate lesson or even a part of a lesson in detail, a more general plan for a month, semester or school year, or just a set of ideas to try out during a lesson. Each lesson plan consists of the similar components, which might slightly vary according to the type of the lesson (Duplass, 2006).

The first component is the objective or the main emphasis on the specific domain, generally defined by state or national standards. It considers the prior knowledge and the students' level of competence and provides some directions to set more specific learning goals. The second component is the content that will be addressed to meet the learning goals (as the

same learning goal(s) can be achieved through different content). The third and fourth components depend on the selection of the content. The third component is a list of materials (equipment) to be used and the sequence of strategies.The fourth is represented by procedures that provide learning the content and achieving the learning goal. Finally, the fifth component of a lesson plan are the formative and summative assessment measures to obtain an overview of the effectiveness of selected teaching and learning strategies in achieving the learning goal(s).

In sum, lesson plans help to produce lessons with unified structure (Jensen, 2001), as they provide teachers with the opportunity to deliberately think about and set the learning goals, select teaching and activities, and materials needed. This is expected in turn to facilitate teaching by making sure that all important components are addressed, but also allowing for time for creativity and personal professional development. A written lesson plan is also useful for developing an objective and accurate assessment, as a lesson plan can provide insights on where the knowledge gaps (or misunderstandings) are most likely to occur or whether particular skills should be fostered.

3.1.3.1 Knowledge and skills needed for lesson planning

Content specific and general pedagogical and conceptual knowledge are prerequisites to developing any lesson plan. Teacher should know the content they teach, what they need to emphasize and which challenges to pay attention to. They should also be aware of teaching strategies to use, their own strengths and weaknesses in terms of teaching; as well as students´ age, preferences or special needs, their level of competence and experience to enable teachers to apply planned teaching strategies. Teachers should be able to formulate a learning goal(s) and decide on the most appropriate type of assessment to determine the effectiveness of the lesson. There are two constructs in recent research that contribute to the understanding of the role and structure of lesson planning as part of teaching competence: professional vision and

teachers' decision-making process about lesson planning. The following paragraphs will introduce these two constructs and their relation to the lesson planning competency.

3.1.3.2 Professional vision

The professional vision concept describes how pedagogical and concept knowledge is used to notice and interpret core elements of classroom situations (Seidel & Stürmer, 2014). It is therefore considered to be one of the essential components of teaching competence. Qualitative research describes several aspects or stages that define professional vision: describe the situation, explain or interpret it, and predict possible consequences. This three-stage structure was confirmed by Seidel and Stürmer (2014) in their study about measuring the structure of professional vision, and also overlaps with the model of measuring competence as a continuum, suggested by Blömeke and colleagues (2015), namely with situation-specific skills building upon each other (perception, interpretation, decision-making) which in turn predict the performance.

Professional vision can be considered a complex cognitive skill with two subcomponents: noticing and reasoning (Seidel & Stürmer, 2014). But as both require application of knowledge, motivational components and different abilities—ranging from setting and clarifying learning goals, providing support and guiding, creating a learning climate, reason and make judgements about the situation in the classroom, predict the consequences of observed activities—it can also be seen as a broader construct – competency. A study by Lefstein and Snell (2010) also defines professional vision as a combination of skills (rather than a single skill) which involves social skills, sensitivity to the classroom situation, dispositions to notice elements of the lesson and capacities to reason the choice of teaching strategies and activities. The empirical findings of the studies by Seidel & Stürmer (2014) and Lefstein & Snell (2010) support that professional vision should be defined as a competency rather than as a single ability or skill.

At the same time, all measurement models of professional vision mentioned above strongly rely on underlying abilities and cognitive skills, which are directly related to observable performance, rather than knowledge or motivation. This aligns with Blömeke and colleagues' (2015) proposal to measure any professional competence by focusing on situation specific processes, rather than stable traits or assessment of knowledge itself.It also supports the argument that cognitive skills are a central building block of any competence, and supports the measurement model presented in this study.

3.1.3.3 Teachers' decision-making

Decision making processes connect real classroom situations with practical actions undertaken by teachers, and therefore these processes are a possible link to connect theory (pedagogical and conceptual knowledge) with actual teaching. Decision making is considered essential for teaching competence. The decision-making research builds on two main assumptions: (1) teachers are professionals that make judgements and decisions in complex situations based on their thoughts and observations before, during and after the lessons, and (2) teaching behavior is influenced by these judgments and decisions (Shavelson & Stern, 1981; Borko, Shavelson & Stern, 1981).

Classroom situations might be unexpected and require immediate actions from the teacher during a lesson, there is little time to make these decisions, and therefore little time to deeply analyse the situation, recall relevant theory and appropriate strategies. As a result these decisions are mainly grounded in experience teachers already have, developed cognitive schemata, teachers' beliefs and values related to teaching. The judgments and decisions made after the lesson are directly connected to the situations that occurred during the lessons (the unexpected ones as well) and aim to assess the lesson in terms of whether set goals were reached, to analyse weak and strong teaching strategies, and to think about possible alternatives and changes that need to be made for the next lesson. Making decisions about planning is therefore an essential part of lesson planning competency. To make these decisions the teacher

should analyse his/her own expectations prior to the lesson and the way the lesson actually went, notice and recognise core elements of the lesson, come up with explanations for possible failures and predict/envision what the next lesson should look like.

This links decision making about planning not only with lesson planning competency, but also with the professional vision. It should be noted that teachers usually do not have time to make notes during their own lesson for later recall and typically they do not even exactly follow their own plan, as unexpected situations require immediate responses. Unexpected situations do not necessarily mean extreme or dangerous situations, but each answer or comment given by students might slightly change the course of the lesson. To be able to analyse the lesson afterwards, a teacher should first of all recall all the relevant and important moments in detail (Shavelson & Stern, 1981).

If teachers (pre or in-service) have an opportunity to observe a lesson conducted by a fellow teacher, it can be easier for them to trace what is happening during the lesson. However, this also sets other challenges: they should infer the cognitive processes and thoughts of the others that lead to the decisions they observe during the lesson. Nevertheless, in general, observing and analysing the lessons held by others (as well as own lessons) provides a great amount of information for planning of own future lessons. The processes that are involved in judgements and decisions are similar for analysing one's own lessons and lessons held by others. In other words, the assumptions about how the lesson was planned (setting goals, choosing teaching strategies and expectations of outcomes) can be implied from the lesson as the final product of this planning.

Van Lehn defines cognitive skill acquisition as acquiring the ability to solve problems in a context, where knowledge is more important than physical strengths (Van Lehn, 1996). To acquire this ability the first step is to notice and recognize the situation, an example or an action as relevant, and as a situation, in which the learned principle can be applied, the second step is the matching of parts of the principle with parts of the problem to be solved (mapping),

followed by third step of principle application and its generalization as a fourth step. Applying Van Lehn's model to the context of lesson planning leads to the following steps: the teacher observes/recalls the lesson, identifies or knows in advance the goal(s) set by him-/herself, recognizes important elements (asking students question), connects these elements with theory (pedagogical and content knowledge), makes inferences about their use and decides whether the technique/strategy was effective for reaching the lesson goal(s), and also decides to use this technique/strategy or change it for the other lesson. In this way the teacher develops an understanding of what teaching strategies match to what learning goal(s) and can solve problems (plan lessons), and make decisions about the use of teaching strategies.

Borko, Shavelson and Stern (1981) developed a scheme which included factors contributing to teachers planning decisions. Among these factors were information about the students, instructional task and educational beliefs that led to the estimation of students' aptitudes and instructional decisions. Further factors that specifically influenced instructional decisions were task related strategies, materials and alternatives for strategies and materials, but also instructional constraints and external pressures. Borko and colleagues (1981) recognised the important role of teacher characteristics, their motivation, beliefs, cognitive processes and inferences. In other words, in addition to national standards, objectives and tendencies, teachers' judgements, expectations, decisions, hypotheses in predicting consequences (actual planning) are essential for planning a lesson. To conclude, a measurement scale of lesson planning should assess noticing and recognition skills, ability to set learning goals and make decisions about the effectiveness and possible alternatives for the teaching strategies to be used.

A school curriculum is a general instructional plan strongly connected with state educational standards. It is subject to the development and implementation of modern trends and scientific findings to better meet students' and state needs, although changes in general take some time to be applied, and the most direct way to notice the change is to look at teacher

education in particular domains. The current scale to assess lesson planning competency is developed in the context of physical education at elementary school and should consider several aspects to evaluate the relevance and content of the planning activities.

3.1.3.4 Lesson planning in the domain of physical education

A review by Balz (2008) on physical education at schools in Germany made its way from the concept of learning different sports disciplines and getting ready for competitions ("Das Sportsartenkonzept" of Wolfgang Söll) to education as acquisition of general skills, social competences and positive attitudes to sports and play ("Handlungsfähigkeitskonzept" of Dietrich Kurz), currently going slightly in the direction of finding balance between acquisition of positive attitude and learning skills and techniques essential for different sports disciplines. Nevertheless, the change in the concept of sports education in the last 30 years was rapid enough to create confusion as to what is expected from a physical education teacher. This is the case especially at elementary school level, where according to the Bavarian regulation (Bayerische Staatsregierung, 2008) a teacher only needs general teaching competencies and no specific qualifications as a sports teacher to work at an elementary school.

The focus of the lessons and the learning goals changed, rapid technological development and the rhythm of modern life set new challenges for physical education at school. To meet these challenge, teacher education should focus on developing teacher skills to become flexible and adaptive experts, who are able to act and continuously learn in new situations. Teaching lesson planning is one of the practical solutions to become an expert, as it helps teachers to implement their knowledge and experience, understand and assess the effectiveness of the lesson and change teaching strategies if needed. The domain of physical education serves as an example and data source for the scale development and the empirical study. The findings can be then generalized and used in the future for teacher education in other domains.

In physical education, the learning goals address one or several of the following aspects: cognitive, psychomotor, attitudinal, self-esteem and social competence. Cognitive goals refer to knowing, understanding, and applying. These goals deal with students' knowledge, and its demonstration. The psychomotor goals involve the acquisition of physical skills, accuracy, dexterity and endurance. Attitudinal, self-esteem and social competence goals involve attitudes, values and emotions that students develop and/or experience, and that evolve, change or end as a result of taking part in physical education activities. Although teaching goals may vary, the lesson structure is usually fixed (Froschmeier et al., 2016). The lesson should start with a lead-in into the lesson and warm-up activity for the muscles, the main part is usually a workout, training and other activities to achieve the learning goal; the lesson ends with a cool-down phase that can also include reflection and follow up. The main part is the longest phase and usually most of the teaching strategies are applied during it, but both lead-in and follow up phases support the learning goal and should be thought through as well.

Methodological approaches and teaching strategies in physical education are similar to the ones in other domains. As in other domains, there are several methodological approaches to introduce new information or skills: inductive or explorative approach, when students can be creative and find out some features on their own. This approach is also usually associated with a holistic approach, when a task/exercise is not broken in individual parts, but trained as a whole; the deductive or teacher guided approach (also called analytic-synthetic) infers that the learning goes from parts to the whole. It is often used to train and better coordinate existing skills.

3.1.4 Defining the measure and the scale development

Inspired by Blömeke et al.'s (2015) suggestion, the measurement model for lesson planning competency evolved. It connects already obtained pedagogical and conceptual knowledge in physical education, teaching experience as well as pre-service teachers' beliefs and motivation with their performance through situation related processes. In the framework of

pre-service teacher education, these processes correspond to lesson planning competency. Hence, the main focus of this study is on measuring the following processes: (a) noticing and recognition of lessons' core elements and teaching strategies, (b) matching these core elements to the theoretical framework, (c) assessment of the effectiveness of these core elements followed by suggesting own ideas to refine the core elements, and (d) the actual planning of one's own lesson. The main emphasis is on fostering the development of complex cognitive skills (analysis and planning of classroom activities). Motivation and knowledge elements are fostered in a pre-service teacher education course and are assessed as control variables.

The teacher's ability to plan their own sports lessons effectively (formulate the learning goal(s), select teaching strategies and decide on effectiveness assessment) might strongly depend on their ability to notice and recognise core elements of a sports lesson; this is also based on their prior conceptual knowledge in physical education and pedagogy. On the other hand, the opposite statement might also be true: if teachers are good at planning, they are also good in noticing and recognizing core elements of the lesson. In line with the review by Gaudin and Chalies (2015) this study adopted video materials to activate pre-service teachers´ knowledge and test their lesson planning competency.

3.1.4.1 Aims of the scale development and hypotheses

The goal of the current chapter is to develop a unidimensional scale fitting the assumptions of the Rasch method to measure lesson planning competency in pre-service elementary school teachers in physical education, which would provide information about levels of competency and possible knowledge gaps, which need to be filled in during pre-service teachers´ training. It is assumed that one-dimensional model can capture lesson planning as a single construct with different processes having different difficulty (noticing on the easier side of the scale, analysing and explaining in the middle and suggesting new ideas on the more difficult part of the scale), but building upon each other to define lesson planning competency (Hypothesis 1). Some might argue that the competency might and even should be

represented as a multidimensional construct as the related processes are distinct even though they contribute to the same performance. A two-dimensional model might rely on the fact that lesson planning consists of two related, but yet distinct processes: noticing, recognising and describing given situations on the one hand, and creating, explaining and predicting outcomes for own ideas (planning own sequences, suggesting alternatives) on the other. A three-dimensional solution assumes, similarly to professional vision, description, explanation and prediction as related but still separate processes (Seidel & Stürmer, 2014).

This study adopts the unidimensional view on lesson planning competency. One of the rationales is that lesson planning competency is a relatively narrow construct compared to teaching competence as a whole. Moreover, this study focuses on the initial stages of scale development and it is therefore more important to measure the related skill in an accurate and the reliable way, rather than to prove that the measure is independent of related processes. The main goal of the study was to select appropriate materials and develop a scale representing the complex skill of lesson planning competency. It is assumed that that the items of the scale can be clustered (Hypothesis 2a) and that the specific competency level can be assessed based on item difficulty (Hypothesis 2b).

3.2 Method

3.2.1 Item Response Theory Approach

The Item Response Theory (IRT) suggests a set of methods that allow to design and validate a measurement scale to address the ability of the test-taker. Unlike the Classical Test Theory (CTT), which relies on the idea of a true score (Cronbach, 1951) and estimates the observed scores as some deviation from this true score due to some random measurement errors, IRT relies on the estimation of probability that a test-taker with certain ability will answer the test item correctly when the difficulty of the item is taken under consideration. This IRT approach helps to design a more robust and fair scale and overcome several shortcomings of CTT. First, item statistics in CTT are not invariant characteristics of the item, but depend on

the particular sample it is obtained from. Second, reliability in CTT strongly depends on the concept of parallel tests, but even the same person can never be exactly the same when retaking the same test (and factors influencing this are hard to control, e.g. fatigue, distraction, forgetting, remembering, obtaining new knowledge or skills beyond the test situation, etc.). Table 3.2 presents a short overview of the main features of Item Response and Classical Test Theories.

Table 3.2

Comparing features of IRT and CTT approaches for scale development

Feature	IRT	CTT
Theoretical model	Obtained score depends on characteristics of the item and the ability of a test-taker	Obtained score is deviation from a true score due to measurement error
Score interpretation	Higher ability is higher probability that the test-taker will answer difficult items correctly	Higher ability is directly represented by higher number of correctly answered items
Item statistics	Invariant	Sample-dependent
Analysis of dichotomous data	Yes	No
Control for accidential events (guessing, errors)	Yes	No
Reliability	Items provide enough information for different levels of ability	Items demonstrate test-retest reliability (Chronbach´s Alpha)

The IRT approach suggests the solution, which is invariant in terms of item characteristics, and the probability of answering it correctly depending on the test-taker ability (but not on other characteristics as social background, gender, etc.). It also enables control for guessing and mistakes, so the amount of ability needed to answer the item correctly is independent from accidental events. It is also worth noting that IRT can better deal with dichotomous data than CTT, the Rasch framework (as part of IRT) is developed to work with dichotomous data. Therefore, Rasch and (IRT in general) suggests ways to design a more fair

scale, especially in learning situations, as it can better handle variability in learning situations and does not assume scores to stay stable from test to test, but rather allows knowledge/ability to change, without being a threat to the reliability of the measure. The Rasch framework is particularly suited for the current study as it allows to distinguish items that can be answered by students with higher ability and therefore provide insights into what knowledge/skills might be missing in students with lower ability.

The Rasch framework requires that several basic assumptions are met (Backer, 2001). The first assumption is that, in case the test-taker knows the correct answer, he/she will answer the question correctly. The second assumption is a local independence of items, which means that answering correctly or incorrectly to each individual item should not fully depend on a correct or incorrect answer to another item. In other words, items are allowed to correlate (as they are supposed to do in order to measure the same latent trait), but this correlation should not be too high, and items should be phrased in a way that they can be answered independently from each other. The most common test in IRT to determine local independence of binary data is calculating tetrachoric correlation, which estimates the possible correlation if the data was measured on the continuous scale. Items are considered locally independent if most of the correlations are below .30 (Cohen, 1988). The third assumption is unidimensionality, that is, the scale only measures one latent trait (knowledge, skill, ability, etc.) but not multiple independent or partly related traits. It is usually tested by goodness of fit tests based on the Chi-square distribution. The fourth and final assumption concerns the shape of the item response curve, also known as the item characteristic curve. This curve has an S-shape and illustrates the relationship between the probability of a correct response to an item and the ability scale. For a typical test item, this probability will be small for those test-takers with a low ability, and large for the test takers with a high ability. The probability of a correct response is near zero at the lowest levels of ability and approaches 1 at the highest levels of ability. Each item in a test will have its own item characteristic curve.

The item characteristic curve is composed of two technical characteristics (independent from each other): (1) difficulty of the item, which plays the role of location index, as it describes where the item is located along the ability scale, and (2) discrimination, which describes how well an item differentiates between test-takers having abilities below the item location and those having abilities above the item location. Discrimination reflects the steepness of the item characteristic curve in its middle section, that is, the steeper the curve, the better the item can discriminate. Under the Rasch framework the discrimination parameter is fixed at the value "1" for all items. Only the difficulty parameter can take on different values and due to this the Rasch model is often referred to as the one-parameter logistic model.

3.2.1.1 Scale reliability, validity and fairness

In the IRT it is not meaningful to compute internal consistency coefficients, because there is no single standard error of measurement (and thus measurement precision). IRT conceptualizes test precision as "information" on the trait level being measured and the ability to distinguish between two respondents. In other words, a reliable test provides sufficient information to distinguish between people with different ability, which means it has enough items with known difficulty parameters in the range where most of the respondents are situated. The EAP and WLE reliability coefficients are typically used to determine the *reliability* of an IRT scale. EAP is the average proportion of the uncertainty in the location of each student. WLE measures the proportion of uncertainty in the location of each item. In other words, the scale should have enough items to distinguish between students in terms of their ability and enough students to distinguish between items in terms of their difficulty. Both coefficients should be above .75 for good reliability, values above .65 are considered acceptable (Baker, 2001).

Model fit is the main *validity* measure of an IRT scale. To identify model fit, Chi-square statistics are typically used with significant p-values indicating model violation, i.e. misfit. Andersen's Likelihood-Ratio-Test (LRT) allows for an assessment of the assumption

that item parameter estimates do not differ across the subsamples but for random variation by comparing the conditional likelihood of the entire dataset. It compares the goodness-of-fit of two models, an unconstrained model with all parameters free and its corresponding model constrained by the null hypothesis to fewer parameters, to determine which model offers a better fit for the sample data. In case of Rasch model fit, it estimates whether the 1-parameter exponential distribution (only difficulty parameter is allowed to vary) is significantly different from the unconstrained 2-parameter exponential distribution (both difficulty parameter and discrimination coefficient are allowed to vary). If the LRT p-value is less than the alpha level (usually 0.05), it can be concluded that the unconstrained 2-parameter model offers significantly better goodness-of-fit than the 1-parameter model for the data.

There is also a set of assumptions about the *fairness* of the items in the Rasch model. A fair item is an item that contributes to assessing the ability of the test-taker, and should not be affected by other factors such as gender, social background, country of origin, etc. If the assumption of fairness is true, items do not have an unexpected item characteristic curve and the probability to answer the item correctly depends only on ability. The Wald test is used to check this assumption: the item responses are split into two groups (mean or median split, alternatively gender or any other dichotomous variable) and subsequently it is determined whether the item has a similar item pattern for both groups. Significant p-values indicate violation of the assumption: if patterns are different for low and high ability test-takers (male and female, etc.) there might be a factor other than ability, explaining the difference (e.g.., being a non-native speaker and not understanding the items, or understanding the item differently due to educational or social background), which in turn defines the item as unfair. Unfair items must either be removed or rephrased (and retested) prior to use of the scale.

3.2.2 Sample and Design

To develop a measurement instrument that aims at assessing lesson planning competency (noticing and recognition of core elements of a sports lesson by pre-service

teachers as well as own lesson planning skills), two independent data collections were conducted (2016: n = 94; 2017: n = 84; overall N = 178 participants) during the annual "Basic Qualification Sports" course at a German University. The 2016 sample consisted of 94 pre-service teachers from the elementary school education track (73%) and the special education track (27%). They were mainly female (97%), on average enrolled in the 6th semester of their studies (SD = 1.14) and their mean age was 22.7 (SD = 2.3). The 2017 sample consisted of 84 pre-service teachers from the elementary school education track (61%) and the special education track (37%); 2% did not specify their track. They were mainly female (97%), on average enrolled in the 6th semester of their studies (SD = 1.78) and their mean was 24 years (SD = 4.1). The same data-collection procedure was applied in 2016 and 2017 to ensure comparability.

The main objective of the course was to prepare pre-service elementary and special needs school teachers to plan and hold their own sports lessons. Therefore, the participants received a 5-day course on teaching methods in physical education (for a detailed overview of the 2017 course in German, please see the Appendix II). The 5-day course consisted of a lecture at the beginning followed by several sessions of observations of examples of sports lessons in elementary school. The lecture provided pre-service teachers with theory and practical implementation of different strategies to teach sports to elementary school children. Afterwards, the participants observed (and had the opportunity to participate in the role of pupils) up to 8 modelled sports lessons per day. They also had the opportunity to observe one or two authentic lessons with elementary school children held by an experienced teacher to obtain more insight into how the modelled lessons would play out in authentic settings and what potential challenges might emerge. To support learning in such an intensive learning environment the pre-service teachers received learning diaries and were asked to write down their observations, thoughts, criticism and suggestions about the lessons they observed.

3.2.3 Lesson Planning Competency: Test Tasks and Scoring

In order to measure lesson planning competency, two types of assessment tasks based on video clips were used: multiple choice questions and open ended questions. Multiple choice questions were focused on pre-service teachers' noticing/analytical skills and open-ended questions assessed the pre-service teachers´ analytical and planning skills. Both tasks contributed to the creation of the single lesson planning competency scale. However the difficulty of the items varied from relatively easy items of multiple choice questions, more difficult items in identifying learning goal and teaching strategies to the most difficult items of the planning task. Multiple dichotomous items coming from both assessment tasks were aggregated using the Rasch framework into a unidimensional scale. Video clips were used as input for the assessment tasks, because they, allowed to present authentic classroom situation to pre-service teachers and also enabled the scale developers to have maximal control over what is shown to the pre-service teachers, to carefully watch and discuss the showed clip and to work out the expert solution for all the questions.

The two video clips were selected from a collection of sports lessons video examples from a book on physical education for elementary school (Froschmeier et. al, 2016). The criteria for this selection were following. First, since the instrument would be used to test German-speaking elementary school pre-service teachers, the video clips of classroom sequences should be recorded in the German language and in the German elementary school context. Second, the video clips should be recorded during physical education lessons. Third, the video clips should activate pre-service teacher's prior knowledge so the clips should be stimulating and activating, but at the same time not too complex and/or distracting. Fourth, it was important to select two video clips that provide enough information to the all components of a lesson planning. Fifth, these video clips needed to be different enough (on a superficial level) to avoid limitations of using the same material twice: memory effects, decrease of motivation, etc. At the same time, video clips had to be similar in terms of content, level of

competency and knowledge required to be comparable enough, so that they could be used for pre- and post- test during the study

The selected videos represented the common view on how physical education lessons are enacted, without trying to select especially good or bad examples. Therefore, the videos were relevant for the pre-service teachers and they offered the opportunity to notice strengths and weaknesses. Table 3.3 summarizes the main features of the two video clips.

Table 3.3

Characteristics of video clips

	Video 1	Video 2
Topic	"Drum Dance"	"Butterflies"
Music	Voice + drumming	Recorder + CD
Equipment	Gymnastic ball, drum sticks, hold	Colourful juggling veils, cones as borders, music
Class	Small group of 5 pupils	Whole class (ca. 25 pupils)
Goal (formulated by expert teacher)	Pupils develop their coordination and sense of rhythm by repeating the demonstrating moves	Pupils develop their coordination and sense of rhythm by repeating and inventing new moves
Length	3 min	
Emphasis	Coordination, repeating the movements	
Age	3rd class elementary school (age 8-9 years)	
Teacher	Same teacher	
Structure	Exercise from the main part of the lesson, switching learned and free movement phases	

3.2.3.1 Assessment task to measure noticing/analytical skills

As part of the lesson planning competency noticing and analytical skills were assessed with a set of nine multiple-choice questions with multiple correct answers and three open ended questions based on video observation (identifying learning goals, strong and weak teaching strategies). Both the multiple choice questions and open-ended questions were designed to be used with (at least two) different videos, therefore correct answers might vary depending on the video used. Experienced physical education teachers and co-organisers of the course ($N = 3$) were asked to watch the two selected video clips and answer in advance (a week before the course started) two sets of questions: 1) the multiple-choice and 2) the open-ended questions. These expert teachers reached the perfect agreement on the both sets of questions for the both video clips. In this way, the expert solution was established.

3.2.3.1.1 Multiple-choice questions set

The set of nine multiple-choice questions was developed to address the dimension of recognition of the core elements of the lesson. This skill is based on attentive observation and the conceptual knowledge the pre-service teachers acquired in their studies and the answers strongly depend on the video observed as they relate to its elements. It was hypothesised that these items would be situated on the easier side of the scale, as they address the attention of the participants and some basic knowledge (equipment used, space needed). The multiple choice questions (questions 2 to 10 in the questionnaire in Appendix IV and the coding manual in digital Appendix) addressed the part of the lesson observed, the main goal of this lesson part, use of the sports gym space, providing opportunities to move for pupils, use of teaching methods, providing safety, assessing complexity of the exercises depending on pupils age, assessing opportunity of pupils to be creative and their emotional response to the classroom situation.

For multiple choice questions several answers were possible and the options were designed in a way they can be treated as independent (choosing one option did not

automatically eliminate the other option or give a hint for another question). The pre-service teachers' answers to multiple choice questions were coded as either agreement (coded as 1) or disagreement (coded as 0) with the expert solution. This allowed the pre-service teachers to get a score not only for choosing the correct answers but also for not choosing the incorrect answers (see the coding manual in the digital Appendix for expert solutions).

3.2.3.1.2 Open-ended questions set

The second set of questions (open-ended questions) was developed to address more difficult parts of the scale, as the items required use of prior pedagogic and concept knowledge about lesson structure and objectives to identify learning goals and strategies, which was only to some extend observable during the video clip. The open ended questions (task 1-2 in the manual) were to 1) list up to three learning goals the teacher might have formulated for the exercise you have observed, and 2) list up to 3 strategies that you consider as effective to reach the learning goals stated above and up to 3 that you consider rather weak. Pre-service teachers were also asked to shortly explain their decisions. According to the expert solution, it was coded if pre-service teachers mentioned the goal from the expert solution (coded as 1) or not (coded as 0) independently for up to three listed goals. Additionally, relevance to the observed lesson, elaboration, use of professional language, and several characteristics of goal formulation were coded. In the goal formulation, mentioning student or teacher activity, the measure to assess if the goal is achieved and mentioning future orientation of the goals were coded independently for each goal mentioned. For the weak and strong strategies, relevancy, elaboration, use of professional language, mentioning attitudes to the observed teaching strategies and mentioning possible alternatives for weak strategies were coded. All the variables were also coded as 1 if they were mentioned and as 0 if not. Codes for the open-ended questions were also assigned in a way to allow considering these items as independent. The detailed rules are presented in the coding manual (see the digital Appendix).

3.2.3.2 Assessment task to measure planning skill

The items of the planning task were designed to address the pre-service teachers' planning skills, i.e. formulating own ideas about the lesson. The task (task 3 in the manual) was asking pre-service teachers to write how would they continue the lesson on the video, what would be the next steps and strategies. They were asked to think about possible Lead-in, Follow up, other exercises to foster development of skills and competencies and achieving the learning goals they previously identified for the observed lesson. Answers to the open ended question concerning planning were also coded (as 1 or as 0) for relevancy, elaboration, use of professional language, suggesting alternatives for exercises or equipment, mentioning learning goals and presenting unique ideas as well as for being realistic (in terms of their implementation in the context of physical education in elementary school). The coding rules allowed to code different items independently from each other. The detailed rules are presented in the coding manual (see the digital Appendix).

3.2.3.3 Coder training and reliability

To validate and improve the quality of the coding manual, the author and one more independent coder used ca. 10% of the pre-service teachers answers to open ended questions from both assessment tasks (N = 20) from both 2016 and 2017 data collections for a coder training. To consider both video clips used for pre- and post-tests, data from both data collections was combined and selected in the way that both videos were presented as pre and post-test. Coders (author and the assistant) were asked to go through the coding manual in advance and ask questions if they occur. In the first round of independent coding the agreement was between 90% and 100%; the codes for each problematic item were discussed and the manual was slightly refined with some more examples and more precise coding rules. For the second round 20 additional answers were coded by the author and the assistant. The coders reached perfect agreement for each item of the scale. The rest of the tests were coded solely by the author.

3.2.4 Procedure

The courses and data collections in 2016 and 2017 were similar in terms of the procedure (same theoretical input, observation of modelled lesson, use of learning diaries, pre and post-test), but differed in terms of additional instructional support provided during the treatment phase (observation notes with and without scaffolds) and the delayed post-test, which took place only in 2017. Therefore, in this chapter the common part of the procedure is elaborated. The specific procedure for 2017 is elaborated in the next chapter (See Chapter 4). Table 3.4 provides an overview over the procedure in both 2016 and 2017.

Table 3.4

Procedure of collecting data in 2016 and 2017

Activity	2016	2017
Theory input	Book chapter + lecture	Book chapter + lecture
Learning diaries	provided	provided
Assignment to experimental/control cond.	no	yes
Pre-test	Beginning of day 1 Video 2	Beginning of day 1 Video 1
Observation of modelled lesson with elementary school children	Day 2	Day 1 and Day 4
Treatment	No additional treatment	Fostering formulation of learning goals
Post-test	Day 4 Video 1	Day 1 Video 2
Delayed post-test	no	Day 4

Introduction. At the start of the basic qualification course, the pre-service teachers were informed about the content and organization of the course and the data collection. Participants signed an informed consent and received a copy for further reference. Participants were asked to use a personal code (consisting of a combination of two letters and two digits), that allowed the researchers to associate all test results from the same person and subsequently anonymize the answers. Participants were also informed that participation in the data collection was voluntary and would not have an effect on their grades or course completion.

Pre and post-test. A short instruction (exactly the same for both pre- and post-tests) was provided: the pre-service teachers had a few minutes to become acquainted with the questions before they viewed the video. The video lasted for 3 minutes and the pre-service teachers had 12 minutes to answer all the questions.

3.2.5 Statistical Analysis: Scale development

To measure lesson planning competency two scales (one for each video clip) using information from both pre and post-tests of two data collections were developed. The scales included dichotomous items from answering multiple choice questions (MCQ: 40 items), open-ended questions on identifying learning goals (OQ: 24 items) and open questions in teaching strategies (OQ: 27 items) and the open-ended planning task (OQ: 7 items) and four more items for the equipment noticed during video observation, in total 102 items for each scale were scored initially.

Scale development and analysis was performed in R with the “eRm" and "ltm" packages (Mair, Hatzinger, Maier, & Rusch, 2016; Rizopoulos, 2017). To obtain fit indices (Andersens Likelihood-Ratio-Test and Wald test using mean split), unidimensionality estimates, item and scale characteristics plots, the "TAM" package was used (Robitzsch, Kiefer, & Wu, 2017). Finally, for item difficulty and person ability estimates, as well as EAP and WLE reliability coefficients, the "WrightMap" package was used for plotting item

difficulty and personal ability (Irribarra & Freund, 2016). Please see digital Appendix for the R code.

3.2.5.1 Item-Selection Algorithm

As mentioned in the section 3.2.1., there are a set of assumptions to follow for creating a Rasch scale. To meet these assumptions a set of rules was created to further select items eligible for the scale measuring lesson planning competency. Within the preparation to the analysis some items were excluded from the analysis: 1) items with no variance providing no information about the competency; 2) items providing irrelevant information to the measured competency (for example writing a personal attitude or emotional reaction to a strategy did not contribute to analytical skills at least in the framework set in this chapter). Some items needed to be recoded due to practical issues.

The first two open-ended tasks on identifying learning goals and teaching strategies required students to list up to three learning goals/strategies, which created some difficulties, as some students mentioned one, some mentioned two or three. As the researchers were more interested in the quality of the answers written, than in their exact amount, the variables for the first two open-ended tasks were recoded. To maintain the dichotomous coding three variables were created instead of coding items separately for each learning goal/strategy: (1) “mentioned in at least one learning goal/strategy out of three”, (2) mentioned in at least two learning goals/strategies out of three, and (3) mentioned in at least three learning goals/strategies out of three. The new variables had same names (e.g. “Elaboration”) and number to indicate amount of mentions (1, 2 or 3). This recoding in turn created a threat to the assumption of local independency of items. This was addressed by including only one of the three variables in the scale as one of the item selection step.

To make the selection and identify problematic items (i.e., items with very low variance, items with unexpected characteristic curve, potentially unfair items, etc.) the decision making algorithm about item inclusion was developed and applied in automated item-selection

in R. To be included into the scale item should (1) not be highly correlated with another variable due to recoding procedure, (2) have insignificant p-values in Wald test (above .05), and (3) if added, contribute to higher EAP and WLE reliability (both indexes should be .65 or above).

3.2.5.2 Procedure of Item Selection

The procedure of the item selection was based on the rules identified in the previous section and was performed automatically by running an R-script. This section describes the procedure step by step and includes some explanations of the steps. At the first step irrelevant items (attitudes, focus on teacher activity in goal formulation) and items with no variance were deleted, which resulted in scale consisting of 84 items for video clip 1, and 86 items for video clip 2. The items were regarded as irrelevant as answering these items was not directly connected with the lesson planning.

Attitudes and emotional expressions about the weak and strong strategies used during the video were considered rather a personal preference and could not be interpreted unambiguously in terms of lesson planning competency. Therefore, the attitude variable was excluded from the analysis for every weak and strong strategy mentioned. Mentioning student and teaching activity was coded as two separate variables initially to assess the quality of goal formulation. However, mentioning the activity of the teacher does not contribute to the formulation of a good learning goal. Using the reversed item would also lead to ambiguity in interpretation. Therefore, the item was excluded from the analysis for each of the formulated goals. Table 3.5 provides more information on deleted items for both video clips.

Table 3.5.

Irrelevant items and items with zero variance deleted from the analysis

Reason of deletion	N for Video 1	N for Video 2	Comment
Irrelevant item: OQ_learning goal	3	3	3 items were same
Irrelevant item: OQ_strategies	6	6	6 items were same
No variance: OQ_strategies	5	2	0 items were same
No variance: MCQ	0	1	0 items were same
No variance: equipment_used:	2	2	0 items were same
No variance: OQ_learning goal	2	2	2 items were same
Total excluded	18	16	11 items were same

The second step was to calculate the distribution of item difficulty to assess the amount of items of different difficulties included in the scale. The Figures 3.1 and 3.2 show the initial difficulty distribution of the included items for two video clips after deleting irrelevant items and items with zero variance. Initial difficulty distribution was used to select one item among dependent items to maintain the most information about different levels of lesson planning competency.

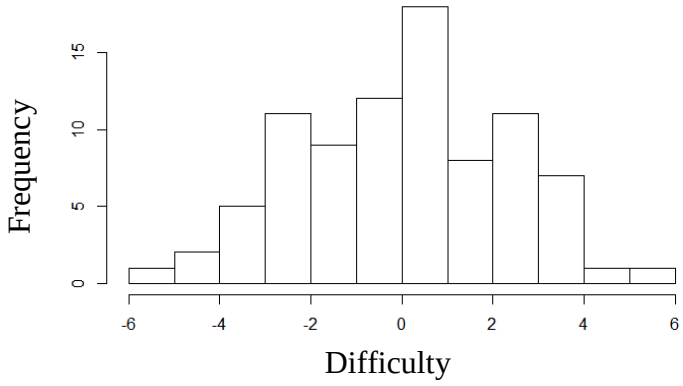


Figure 3.1 Item difficulty distribution for video clip 1 (84 items)

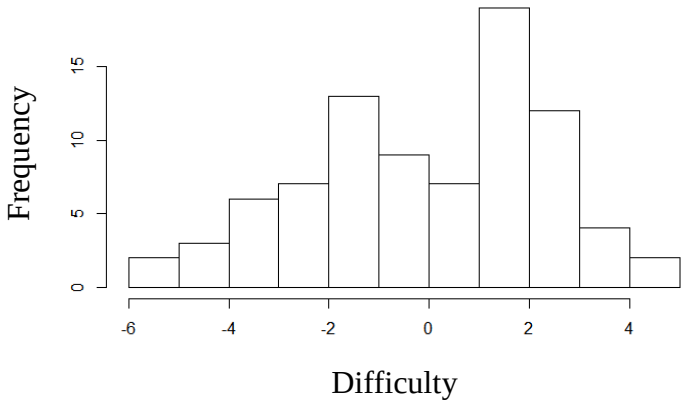


Figure 3.2 Item difficulty distribution for video clip 2 (86 items)

The third step was checking the Rasch assumptions for model fit. The initial scale for video clip 1 violated model fit according to Andersens Likelihood-Ratio-Test: $\chi^2(70) = 98.02$, $p = 0.015$, which indicates misfit of the 1-parameter (Rasch) model, as the item parameters in the subsample and the whole sample differ more than by random variation, which might in turn be due to items which are not locally independent, but also inadequate or unfair items. Video clip 2 does not violate this assumption: $\chi^2(73) = 78.40$, $p = 0.312$, but it still includes items which are not locally independent due to coding procedure, and might also include other problematic items, so further check and item selection were made.

The fourth step was step-by-step removal of items with heterogeneous standard error. Eligible items should have similar discriminatory ability. Heterogeneous standard error indicate a different discriminatory ability for some items, which violates the Rasch assumption, and subsequently might indicate the items that are more or less sensitive to reflect the level ability needed to correctly solve the task. Lack of sensitivity to reflect the competency level indicates problems related to standard error: (1) low variance due to too difficult or too easy items that do not differentiate between high and low competency, which might indicate a ceiling effect, and (2) extremely high variance, which indicates that there are other factors in addition to the level of competency responsible for high variation between difficulty scores. Problematic items violate the assumptions of Rasch model and hinder the interpretation of the competency levels.

The fifth was step-by-step removal of items with significant Wald test. The R-script was programmed to use the mean split to evaluate whether the answering pattern was similar for low and high competency groups. Only items with similar patterns were considered fair and eligible for the selection. Similar patterns indicated that items assess the desired competency, but not other personality, background or environmental factors that are out of control during the intervention.

The sixth step was ensuring Local Independency of items. To exclude dependent items but keep the scale as informative as possible the selection was done in reference to the difficulty distribution histogram. If more than one of the dependent items has been left in the scale after previous steps, the one that provided the item difficulty at the competency level with little information (i.e. there were little items with similar difficulty) was left in the analysis to maintain the amount of information at this competency level. Whereas if an item shared similar difficulty with many other items in the scale, it was removed because there was enough information at this level of competency from other items. This selection procedure allowed to avoid the scale to fail to discriminate between higher and lower competency level of pre-service teachers and maintain the high reliability.

3.3 Results

3.3.1 Description of the final scales

3.3.1.1 Items in the scale and difficulty distribution

After applying the item-selection algorithm, the final scales included 48 items for video clip 1 and 50 items for video clip 2. As hypothesised, the multiple choice questions (MCQ) and listing of the equipment used (OQ_equipment used) provided items for the easier part of the scale, open-ended questions about identification of learning goal (OQ_learning goal) and teaching strategies (OQ_strategy) provided items of intermediate difficulty and the open-ended questions addressing planning of own lesson (OQ_planning) provided items on the most difficult side of the scale. Table 3.6 provides information on the remained items in the final

scales. The other items were taken out from the analysis during selection procedures because of violating the Rasch assumptions of local independency (20 items from video 1 and 18 items from video 2), unidimensionality, homogeneity of discrimination coefficient and fairness (11 items from each of the videos). The most problematic items originated from the open questions, several items from multiple choice questions were taken out from the analysis because of low variability (5 items from video 1 and 7 items from video 2).

Table 3.6

Set of final items in the lesson planning competency scales

Items	N for Video 1	N for Video 2	Comment
MCQ	35	33	28 items were same
OQ_learning goal	3	4	2 items were same
OQ_strategy	4	5	4 items were same
OQ_planning	4	6	4 items were same
OQ_equipment used	2	2	1 item was same
Total	48	50	39 items were same

Figures 3.3 and 3.4 present the difficulty distribution of the final item-set for each of the two video clips. The two scales had similar amounts of items and the difficulties of the single items were similar for the both video clips.

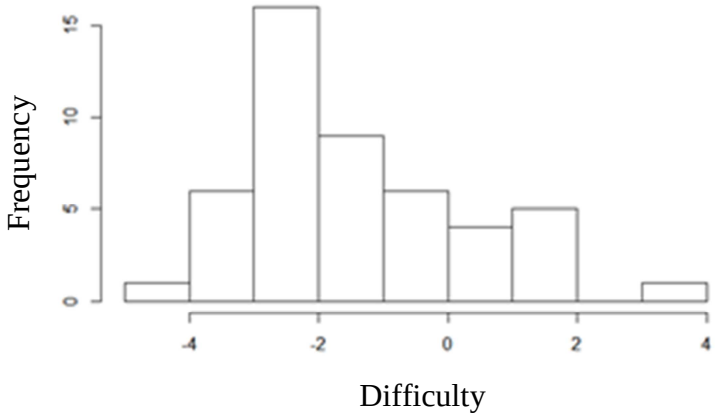


Figure 3.3 Final item difficulty distribution for video clip 1

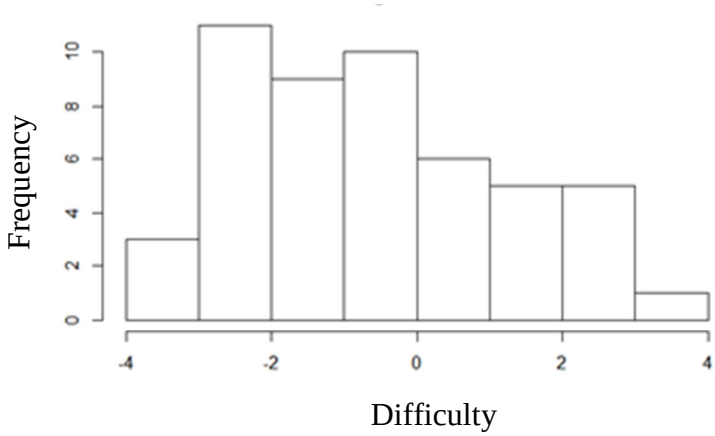


Figure 3.4 Final item difficulty distribution for video clip 2

The two scales based on pre-service teachers´ answering of the final set of the items had similar scale characteristic curves, suggesting that they were comparable and could be used as pre and post-tests to assess the lesson planning competency. Both scales also provided more information on lower lesson planning competency level (left skewed), which made them suitable for pre-service teachers or teachers with little experience in lesson planning. Figure 3.5 shows the scale characteristic curves for two video clips.

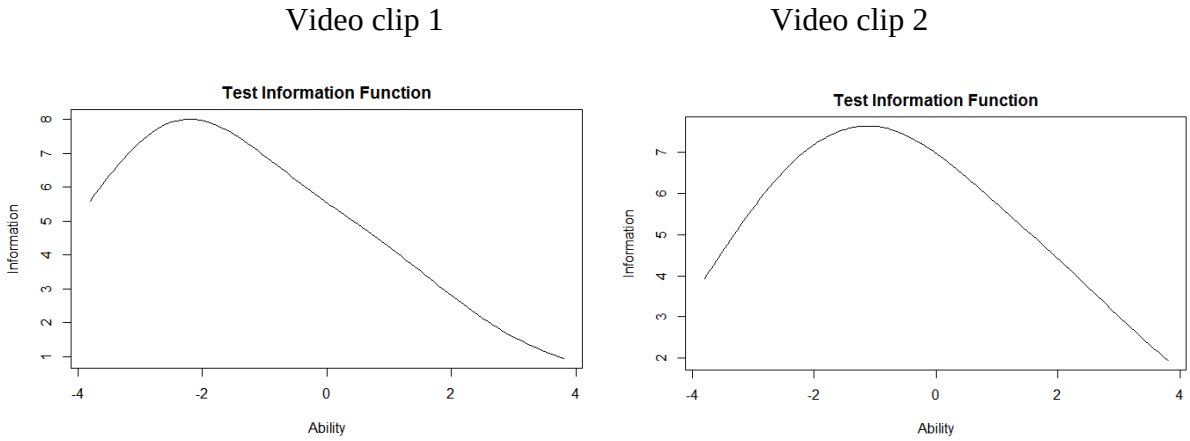


Figure 3.5 Scale characteristic curves for video clip 1 and video clip 2.

3.3.1.2 Model fit and reliability

The final scales for both video clips had non-significant p-values for Andersen's Likelihood-Ratio-Test: (Video 1: p = 0.74; Video 2: p = 0.08), indicating a good fit of the one

parameter model. Both final scales also had an acceptable level of reliability: Video 1 WLE Reliability = 0.667, EAP Reliability = 0.692; Video 2 WLE Reliability = 0.697, EAP Reliability = 0.698. Hence, both scales provided enough information to distinguish between different levels of lesson planning competency in pre-service teachers.

3.3.2 Standardization of the two scales

To enable the assessment of gain in lesson planning competency, the competency scale for both video clips was standardised for pre-tests by extracting the mean value from the individual lesson planning competency score.

3.3.3 Interpretation of ability scores

The Rasch scale allows interpreting the lesson planning competency of individual pre-service teachers in terms of specific items that the pre-service teacher will probably answer correctly while being at a certain level of lesson planning competency. Table 3.7 depicts the items in each of the two scales ranked according to their difficulty.

Table 3.7

Item difficulties for video clips

Video 1 “Drum dance”

Item clusters	Competency level required to answer the item correctly
OQ_planning: Alternatives OQ_planning: Prof. language OQ_planning: Own ideas OQ_planning: Elaboration OQ_strategy: Notice weak MCQ: Security specific	High (2 - 4)
OQ_strategy: Notice effective OQ_strategy: Prof. language OQ_learning goal: Stud. activity OQ_strategy: Elaboration OQ_equipment used: music	Medium (-0.5 – 0.5)
MCQ: Complexity MCQ: Creativity MCQ: Equipment general MCQ: Lesson part MCQ: Lesson objective MCQ: Opportunity to move MCQ: Relevant goal MCQ: Security general MCQ: Student reactions MCQ: Teaching strategy	Low (-2 - -1)

Table 3.7

Item difficulties for video clips (continued)

Video 2 “Butterflies”

Item clusters	Competency level required to answer the item correctly
OQ_planning: Alternatives OQ_planning: Prof. language OQ_planning: Own ideas OQ_planning: Elaboration OQ_strategy: Notice weak MCQ: Security specific MCQ: Equipment specific	High (2 - 4)
OQ_strategy: Notice effective OQ_strategy: Prof. language OQ_learning goal: Stud. activity OQ_strategy: Elaboration	Medium (-0.5 – 0.5)
MCQ: Complexity MCQ: Creativity MCQ: Equipment general MCQ: Lesson part MCQ: Lesson objective MCQ: Opportunity to move MCQ: Relevant goal MCQ: Security general MCQ: Student reactions MCQ: Teaching strategy OQ_equipment used: music	Low ability (-2 - -1)

A pre-service teacher with the *competency of -2* in lesson planning would be able to correctly answer the items about the video that are easily observable but would fail to connect to theory, elaborate and use professional language in the description, as well as fail to define a learning goal and develop a planning task to support it. A pre-service teacher with the competency of 0 would be able to correctly answer items with a difficulty level below 0, more specifically to identify some elements of the lesson (lesson type, main equipment and space needed) but will fail to elaborate and use professional terminology in the description and

planning task. A pre-service teacher with the competency of 2 would answer most of the items correctly and be able to develop the planning task, though s/he still might fail to suggest alternatives to weak strategies or unique ideas during planning or disregard some specific detail in observation or planning.

3.4 Conclusions

The two video clips provided sufficient information about the level of lesson planning competency required to correctly answer the items in the test. The video clips are statistically comparable in terms of difficulty (no significant difference in mean difficulty when both tests are used as pre-tests) and item distribution (similar items have similar difficulty). Different types of items, as hypothesized, tended to indicate different difficulty levels: items on noticing the elements in the video tended to land on the easier side of the scale while planning activities together with use of professional language, connecting the observed fragment to theory and elaboration on strengths and weaknesses have shown higher difficulty. At the same time, all the items provided information about the single construct (lesson planning competency) with acceptable reliability that also supports the theoretical expectations. Therefore, the scale to assess the effectiveness of the instructional methods used during the treatment phasecan be used in empirical research for measuring lesson planning competency in participants at several time points.

Further use of the scale in the empirical research should however acknowledge some limitations. The scale might fail to detect the gain in the lesson planning competency if applied to measure competency of experienced teachers, as the scale provides relatively a low amount of items on the more difficult side of the scale. The scale is more reliable at the lower levels of competency. The unequal distribution of items with high and low difficulty had led to the acceptable, but relatively low scale reliability from the Rasch framework perspective. To further develop the scale and overcome the existing limitation more items to discriminatebetween participants with average and high competency is needed.

4 Fostering Lesson Planning Competency in Pre-Service Teachers

The fourth Chapter presents the empirical study aimed at assessing the effect of the scaffolding within the observational learning on the development of lesson planning competency in pre-service teachers in the domain of physical education. This study was designed based on the findings from the meta-analysis (Chapter 2) and aimed at contributing to the body of empirical research in several ways. First, it focused on lesson planning competency as a complex construct, rather than on the acquisition of a single principle or learning a single technique or didactic principle. Second, it had a relatively big sample compared to the empirical studies in the meta-analysis and implemented a pre-post, control group design with random assignment to control and experimental group. Third, it adopted design features that were underrepresented based on the meta-analysis (in vivo observation, scaffolding during observation).

The empirical study suggested and tested the procedure of scaffolding of learning goal formulation during observation of example lessons in physical education. This procedure is described in detail in Section 4.2.2. It was decided to scaffold the process of learning goal formulation, as it is the initial and the most important step in the lesson planning, and all other steps and decisions rely on it. The study implemented the scale developed to assess lesson planning competency (see Chapter 3) as a single construct. Although the main focus of the study was on the knowledge and cognitive skills underlying the lesson planning competency, the pre-service teachers' beliefs about what is needed for good planning were also taken under consideration and were supposed to represent the motivational component of the competency. Background factors (semester of study, teaching experience, educational track) were also recorded and included into analysis to control for possible moderating effects. During the analysis the researcher formulated additional research questions considering the adherence to the instruction (in the scaffolding tasks) as the predictor of gain in lesson planning competency.

4.1 Theoretical framework and Research Questions

As conceptualized in Chapter 3, teaching competence is a product of (1) pedagogic and domain specific conceptual and procedural knowledge, (2) complex cognitive skills and (3) motivational and affective factors that lead to effective planning and successful performance in classroom situations. Competence development and successful performance require constant learning to maintain up to date knowledge and resolve misconceptions in pedagogic and domain specific knowledge; practicing and developing cognitive skills to implement the knowledge into practice; but also having realistic beliefs about the important core structure and teaching tasks to assure that knowledge and skills are applied in an effective way. Teaching beliefs play a big role in competence development and should be considered not only in planning and designing teacher programs, but also in performance assessment.

The dissertation focused on lesson planning competency as an essential part of teaching competence for several reasons. First, lesson planning competency is based on the tasks that the teacher performs, rather than on the personal characteristics of the teacher. Second, assessing lesson planning allows to trace the processes of decision making and to identify if any of the competency´s core elements (knowledge, cognitive skills or motivational factors) contain misconceptions or need to be fostered. Third, an actual lesson and a lesson plan have similar structure and improving planning competency should contribute to better teaching. It also means that the assumptions about how the lesson was planned (setting goals, choosing teaching strategies and expectations of outcomes) can be implied from the lesson as the final product of this planning. This makes it possible to use observational learning approach to foster the development of the competency.

If teachers have an opportunity to observe a lesson conducted by a fellow teacher, they can see what is happening during the lesson from a different perspective. However, this also sets challenges: they should infer the cognitive processes, thoughts and the learning goals that lead to the decisions they are observing during the lesson. This might be particularly

challenging for pre-service teachers, who have little experience in teaching their own lessons. Although observing and analysing the lessons held by others (as well as own lessons) provides a lot information, this observation needs to be supported with instructional aids, to minimize misconceptions and distractions and maximize the benefit for the teachers. Scaffolding and the other types of additional instructional support are considered to aid learning, especially in case of complex skills acquisition, which requires to process large amounts of different/new information (Dianovsky & Wink, 2011; Glogger et al., 2009). Additional instructional support helps to reduce the cognitive load and therefore free up more working memory capacity for processing, structuring and integrating new information (Kirschner, Sweller, & Clark, 2006), which in turn can enhance learning. Moreover, additional instructional support can be used to support different stages of cognitive skill acquisition as described by Van Lehn (1996).

4.1.1 Research questions

RQ1: To what extend does scaffolding (facilitating the formulation of learning goals) during observational learning, impact the pre-service teachers´ lesson planning competency?

As cognitive skills are an important element of any competence or competency (Stoof et al., 2002) proper instructional support should be used to foster their development. In line with the findings from the meta-analysis (see Chapter 2), it was hypothesized that scaffolding the formulation of learning goals would have a positive effect on the lesson planning competency in pre-service teachers: the experimental condition would outperform the control condition during the post-test phase (Hypothesis 1).

RQ2: To what extend do teaching experience and motivational factors (beliefs about the importance of learning goals) predict the post-test lesson planning competency?

Empirical research suggests that prior knowledge and teaching experience can contribute to knowledge and skill acquisition (Blömeke, Gustafsson, & Shavelson, 2015; Renkl & Atkinson, 2003). Hence, it is assumed that prior teaching experience moderates the lesson planning competency in a way that teachers with more prior teaching experience will score

higher in lesson planning competency compared to pre-service teachers with little or no prior teaching experience (Hypothesis 2a). Motivational aspects are also an essential part of the competence model by Blömeke and colleagues (2015). In this respect the value assigned by students to well formulated learning goals as essential part of lesson planning is relevant, because a higher assigned value can lead to more effort spent on learning how to formulate a good learning goal. Therefore, it is hypothesized that the value assigned to the learning goals predicts the post-test score in the lesson planning competency (Hypothesis 2b).

4.2 Method

The following subsections of the method section describe (1) the context and the structure of the "Basic Qualification Sports" course and the participants; (2) the study design and the procedures applied during data collection; (3) learning materials, instruments and measures used for the data collection; (4) statistical analysis procedures used to address the research questions. The subsection on learning materials and measures is split into a pre-test phase, which includes preparation to the data collection, introductory activities and the pre-test; a treatment phase, which describes the materials and manipulation check measures used during the treatment; and the post-test phase, which describes materials and measures used during the immediate and the delayed post-tests.

4.2.1 Context and Participants

The data collection took place in 2017 during the annual course "Basic Qualification Sports" at a large university in the south of Germany. The course aimed at teaching pre-service teachers to analyse and plan sports lessons by applying theoretical, pedagogical and didactical knowledge they already obtained during studies at university. The focus of the course was to prepare pre-service teachers to teach physical education at elementary school, provide them with pedagogical and didactical hints and reveal potential problematic issues (i.e. establishing safety during the lessons, dealing with differences in pupils' ability levels, focusing on teaching social skills) by modelling sports lessons in different sports arts. The Basic

Qualification course lasted one working week (Monday to Friday) and was offered during the semester break, a week before semester had started. Such scheduling, on the one hand, granted presence of most of the participants, and on the other hand, allowed pre-service teachers to intensively work on the content of the course, without being distracted by other courses or activities. During each of the five days, the pre-service teachers had the opportunity to observe and participate in the role of elementary school pupils in up to 8 lessons, modelled by expert teachers. The lessons lasted approximately 30-40 minutes; the rest of the time (5 to 15 minutes) was used for discussing the content of the lesson, teaching strategies and possible alternatives for didactics and used materials. Additionally, pre-service teachers had an opportunity to make notes about their observations (during and/or after the modelled lesson depending on the role they have chosen) using a learning diary. All pre-service teachers were required to make notes about at least two lessons during the day. See Table 4.1 for the schedule of the course and data collection. A more detailed schedule with lesson topics can be found in Appendix II.

The sample of pre-service teachers consisted of 84 pre-service elementary and special education school teachers. Elementary school pre-service teachers made up 61% of the sample, special education pre-service teachers were 37% of the sample, and 2% did not specify their educational track. The pre-service teachers who did not specify their educational track were included in the general analysis, but excluded from the corresponding moderator analysis. Participants were mainly female (97%), which authentically reflects the situation in elementary school teaching. The results of male participants were checked for being significantly different from the results of female participants on all the scales. As no significant difference was found, all the participants were included in the analysis. Participants were enrolled on average in the 6th semester of their studies (SD = 1.78). The mean age of the participants was 24 years (SD = 4.1). The pre-service teachers were randomly assigned to an experimental (N = 43) and a control (N = 41) conditions.

Table 4.1
Time plan for the course Basic qualification Sports and data collection

	Monday	Tuesday	Wednesday	Thursday	Friday
9.15 - 10.00	***9.00 Uhr*** *„Check in", Introduction, Informed consent form, collecting background information.*	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*
10.00 - 10.45	*Warm up lesson*	*Example lesson*	*Example lesson*	Watching modelled lesson with elementary school pupils. DELAYED POSTTEST	*Example lesson*
10.45 - 11.00			**Break**		
11.00 - 11.45	*Lead in lecture PRE-TEST (11.30 – 11.45)*	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*
11.45 - 12.30	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*
12.30 - 13.30			**Break**		
13.30 - 14.15	Watching modelled lesson with elementary school pupils. TREATMENT	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*
14.15 -15.00	Watching modelled lesson with elementary school pupils TREATMENT	*Example lesson*	*Example lesson*	*Example lesson*	
15.00 - 15.15			**Break**		
15.15 - 16.00	*POSTTEST (15.15 – 15.30) Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*	
16.00 - 16.45	*Example lesson*	*Example lesson*	*Example lesson*	*Example lesson*	

4.2.2 Design and Procedure

An experimental pre-post-test-design with a control condition was implemented. Pre-service teachers were randomly distributed to the experimental and the control conditions prior to the start of the data collection. The procedure was split into pre-test phase (included preparation for the course, introduction, collecting background information and the pre-test); treatment phase and post-test phase, which included immediate and delayed post-tests. The overview of the procedure is presented in Table 4.1.

4.2.2.1 Pre-test phase

Introduction. In the beginning of the course pre-service teachers were informed about the objectives and the procedure of the data collection. They were also informed that participation was voluntary and would not affect their grades or course completion. The participants were asked to sign informed consent (see Appendix I) and received a copy for further reference. After collecting informed consents, background information was collected and the participants were assigned to the treatment and control conditions. While collecting background information (see Appendix III) the pre-service teachers were asked to invent a personal code (consisting of a combination of two letters and two digits), and used it in all the forms they filled in during the course. This procedure allowed to associate all tests collected in the run of the course with the correct person, but also to anonymize the data file by using the codes instead of names or other personal data. For the participants to remember the condition they were assigned to and for the researchers to keep the record of the random assignment, the participants were given a corresponding coloured bracelet (orange for control and blue for experimental condition). Pre-service teachers were asked to keep the bracelet for the whole duration of the course.

Lead in lecture. After a brief welcome, introduction and warm up lesson the pre-service teachers received theoretical input (power point presentation) based on the introductory chapter from the book by Froschmeier et al. (2016) which they had read in advance (a week before the

course had started). The lecture also provided pre-service teachers with a brief historical overview of the development of physical education in Germany and the changes in common and specific learning goals set for physical education lessons in elementary school (See Section *3.1.3.4* of Chapter 3 for an overview). Pre-service teachers were introduced to the most up-to-date trends and teaching strategies in physical education for elementary schools as well as state-level requirements and guidelines. The importance of lesson planning and formulating learning goals was explicitly emphasized. The lecture was followed by brief questions and answers about the issues raised in the lecture.

Pre-test. After the lead-in lecture, the pre-test was distributed. A short instruction was given to the pre-service teachers (see Appendix IV for the instructions) and they had a few minutes to inspect the multiple-choice questions and open-ended tasks before they saw the video clip. It was done for two reasons: (1) to focus pre-service teachers´ attention on different aspects of the video clip; (2) to have identical conditions at the pre and post-test phases (since otherwise students would have had an advantage of being familiar with questions on the post-test phase). The pre-test video (a sportss lesson called “Drumming dance”) lasted for 3 minutes and students had 12 minutes to answer the questions (see Section 3.2.3 of Chapter 3 for the questions overview). After the pre-test and before the start of the treatment phase the pre-service teachers participated in an example lesson.

4.2.2.2 Treatment phase

Treatment. The treatment took place after the lunch break. During the treatment phase the pre-service teachers observed a lesson, modelled by the invited teacher and her class of elementary school pupils (age 8-9 years). The invited teacher showed two lessons (60 and 30 minutes long): the first lesson focused on team work and developing social competencies and the second lesson focused on cooperation and ball games skills. The students received structured observation forms (see Appendix VIII-XI). The forms for experimental condition (blue bracelet) and the control condition(orange bracelet) looked very similar at a superficial

glance, but differed in terms of theoretical introduction to formulating learning goals and a special task given to the experimental condition during observation, vs. general information about observation form and no special task received by the control condition. At the end of the treatment phase, the observation forms were collected and scanned. The original forms were returned to the pre-service teachers on the last day of the course (Friday), after the data collection was completed.

4.2.2.3 Post-test phase

Post-test. After the treatment phase was over and pre-service teachers had a 15-minute break the post-test forms were distributed. The very same instruction as for the pre-test was provided (see Appendix IV) and the pre-service teachers were given a few minutes to inspect the questions before they saw the video clip. The post-test video clip (a sportss lesson called "Butterfly Dance") had the same length as the pre-test video (3 minutes). Pre-service teachers had 12 minutes to answer the questions, which were identical to the ones in the pre-test.

Delayed post-test. The delayed post-test was administered two days after the treatment. Instead of a video, the pre-service teachers observed a real sportss lesson (45 minutes, elementary school pupils, 8-9 years old) modelled by another invited teacher and her class. The emphasis of the modelled lesson was on handball techniques. All pre-service teachers received identical observation forms with a brief instruction and structured space for writing down the lesson plan, sketching gym, equipment and/or activity during the observed lesson, and space for further notes. Pre-service teachers were asked to note down all information they considered important for planning their own lesson similar to the one they had observed. The observation forms were collected, scanned and returned to pre-service teachers at the end of the course, so that they had all the notes and documentation from the course for the reference and use in planning own lessons.

4.2.3 Materials, Instruments and Measures

4.2.3.1 Learning materials used in the course

This section provides information about all the learning materials used in the Basic Qualification Sports course, materials and instruments used to measure lesson planning competency and other variables are presented in the following sections. The materials included a book chapter and a presentation, which provided theoretical input for both experimental and control conditions; learning diaries and observation forms, which supported pre-service teachers´ learning from example lessons during the course; and a brief overview of learning goals types provided as additional theoretical input provided within observation forms to the experimental condition only.

Theoretical input, common for all the pre-service teachers consisted of two parts. The first part was an introductory book chapter (Froschmeier et al., 2016) which informed students about the current state of teaching sportss at elementary schools, trends, goals and challenges in physical education. The chapter was sent to students a week before the course started and they were asked to read it before coming to the course. The second part was a presentation, shown during the lead-in lecture, about most essential pedagogical and didactic issues in elementary school physical education. The presentation included discussion of the needs of modern elementary school children, and especially problems, including fewer opportunities to move and play, conceptualizing sports as a competition rather than a chance to support healthy development of children. The presentation also introduced the historical view on physical education and modern concept of sports lessons at schools aimed at meeting needs of the children and the ways it could be adopted by teachers. Additionally, the presentation covered the basics of lesson planning in physical education, specifically the introduction and realisation of the modern sports concept through thorough planning of learning objectives, content, and teaching methods and strategies.

Learning diaries were also used by all pre-service teachers during the course to make notes about the content, equipment, specific exercises and teaching strategies used during the example lessons. The learning diaries were provided in the form of a structured observation booklet in which students had special space to make notes about the structure of observed lessons, exercises and instructions given (an empty table to fill in, divided into lead-in, main part and cool-down phases), draw sketches of gym (empty space in the form of the gym, to note the use of space, activities and the equipment), and another empty page with lines to note down their own thoughts, ideas or critic about the lesson. The learning diaries only served as a support and future reference for pre-service teachers. They were neither collected nor graded or analysed. For an example of the booklet, see Appendix IX.

Observation forms were structured in asimilar way to the learning diaries and used to make notes about the lessons conducted by invited teachers with elementary school children (during treatment and delayed post-test). Unlike learning diaries, observation forms were collected from students and given them back at the end of the course. The front page of the observation forms introduced the lessons´ topics and the rules of using the observation forms. For the experimental condition, the front page also introduced additional theoretical input about how to formulate the learning goal and what types of learning goals exist. Beyond the table for the lesson structure, place for sketch of the gym and place for making own notes (as in the learning diaries), observation form for the experimental condition contained a special task about identifying the learning goal of the observed lesson and ranking observed activities according to how much they supported the learning goal.

4.2.3.2 Instruments and measures

4.2.3.2.1 Pre-test phase

Background information collected during the introduction phase included: age of the participants, gender, educational track (elementary or special education school), current

semester, the amount of credit courses taken (in %), teaching experience (pre-service teachers were asked if they had any prior teaching experience, and if yes, about its type and duration).

Beliefs about the importance of learning goals were measured in a ranking task during the introduction phase of the data collection. Students were provided with five statements and asked to rank them from 1 to 5 depending on how important they thought these statements were for planning the lesson. The first statement was formulated as a tangible measure of students' progress after the lesson, the second mentioned the learning activity student should do during the lesson, the third mentioned the activity teacher would perform during the lesson, the fourth was the statement about the general content of the lesson, that the fifth mentioned the necessity of topic independent warm-up exercise. The expert solution for the ranking was designed according to the theory of lesson planning and learning goal formulation and in cooperation with expert teachers. The difference in ranking made by pre-service teachers and expert solution was added as a covariate in the further analysis.

Lesson planning competency (pre) was measured by the scale developed in Chapter 3, for the pre-test video clip 1 ("Drumming Dance") was used. See section 3.2.3.1 for the description of the items in the scale. The scale reliability for video clip 1 was above .67 (reliability is considered acceptable if above .65).

4.2.3.2.2 Treatment phase: manipulation check

Manipulation check: adherence to instruction. As the course and specifically the treatment phase were structured in a way that gave the experimenter little opportunity to interfere during observation, provide clarifications or control the completion of tasks, a manipulation check was developed. Its aim was to monitor to what extend the students followed the instructions and completed the tasks (identifying the learning goal of the observed lesson and the activities that helped to achieve the learning goal) during the treatment.

The manipulation check consisted of 5 tasks (open-ended questions) for each of the two observed lessons. The first task was to identify the result of the lesson, pre-service teachers

were asked to identify what pupils were able to do better after the end of the observed lesson than before. The second task asked for the observed evidence that the learning goal was achieved. The third task was to identify/nominate at least one of the activities noted during observation as an activity that contributed to achieving the learning goal (productive). The fourth task asked students to mention at least one of the activities as neutral (not contributing to the achievement of the particular learning goal, but also not conflicting with it). The fifth task asked students to mention at least one of the activities during the observed lesson as not contributing to achieving the learning goal, but conflicting with it (counterproductive). Each task was coded as 1 or 0. A score of 1 was assigned if the task was completed, and a 0 was assigned if the task was not completed; the quality of the responses was not coded (see Appendix X for more details on the questions and coding manual). A scale was constructed out of these 10 items to measure, to what extend the students in the experimental condition followed the instructions provided during the observation task and reflected the percentage of instructions followed.

4.2.3.2.3 Post-test phase

Lesson planning competency (post) was measured by the same scale as during pre-test (see Chapter 3 for scale description), for the post-test video clip 2 (“Butterflies”) was used. The scale reliability (video clip 2) was above .69 (reliability is considered acceptable if above .65).

Delayed post-test. The aim of the delayed post-test was to identify if pre-service teachers learned to identify and formulate the learning goals of observed lessons and if the treatment had an effect on the overall amount of detail in notes taken during observation. The delayed post-test was developed to check if the students continued to use the scaffolds that they were introduced to during treatment phase, even without any prompting to do so. The measure consisted of two sets of items, which contributed to the creation of two subscales.

The subscale *“Learning goal”* consisted of 7 items and focused on identifying and formulating the learning goal. The items were similar to the learning goal set of items in open

ended questions of the lesson planning competency scale). More specifically, the items focused on mentioning the learning goal, elaboration on it, using professional language, mentioning student activity in goal formulation, etc. The subscale *"Observation notes"* consisted of 11 items and focused on general issues with respect to the level of detail and quality of notes and sketches made during the observation. Cronbach's alpha for the subscale "learning goal" was .77, whereas for the subscale "observation notes" it was only .50. Hence, only the subscale "Learning goal" was used for further analysis.

4.2.4 Statistical Analysis

To answer the research questions the statistical analysis was run in R using the "stats" package (R Core Team, 2017). To answer the first research question, experimental and control condition scores in lesson planning competency were compared using ANOVA procedures, while repeated measures ANCOVA was implemented to test for the significant gain in the lesson planning competency from pre to post-test. The statistical model included test (pre- or post-), condition (experimental or control) and their interaction. The intercept was forced to be 0, because the pre-test was conditioned to have a mean of 0 (see Section 3.3.2). This could potentially cause heterogeneous slopes for pre- and post- tests, so the model was adjusted to accommodate that. To answer the second and third research questions, multiple regression procedures were implemented. The R-script for the analysis is available from digital Appendix.

4.2.5 Results

4.2.5.1 Preliminary analyses

To ensure that differences between the lesson planning competency of the experimental and control condition were due to the treatment, but not demographic factors, the preliminary analyses were performed. There was no statistically significant difference identified between control and experimental groups in terms of age ($F(1,77) = 0.43$, $p = 0.53$), teaching experience (F (1,79) = 0.87, p = 0.35) and education-related variables: EWS (F (1,73) = 0.66, p = 0.41), educational track (F (1,78) = 1.06, p = 0.31), semester (F (1,78) = 0.62, p = 0.43). Pre-service

teachers in the experimental and control conditions did not differ in their lesson planning competency at the pre-test phase (see Table 4.2).

As the experimenter could not interrupt the observation for clarification questions or additional instructions, a manipulation check was conducted to determine the degree to which the students followed the instructions during observation. The manipulation check indicated that only one out of 43 pre-service teachers in the experimental condition followed all instructions and completed the entire task (see section 4.2.3.2.2 for details) during observation of modelled lesson; fifteen out of those 43 pre-service teachers completed 50% or less of the observation tasks during observation. On average the pre-service teachers completed 59% of the task during observation. It is important to mention that the measure did not include the content or quality of the answers provided, but only whether pre-service teachers answered the questions. The slightly above average percentage signals that a large portion (30%) of students in the experimental condition failed to follow the instruction. As a result this introduces high variation within the experimental condition, which should be taken into consideration when interpreting the effect of the treatment on students' gain in lesson planning competency. Therefore an additional post-hoc research question was set:

RQ3: To what extend does adherence to instructions during the treatment phase predict the lesson planning competency at the post test phase? It is assumed that adherence to the instructions will be a significant positive predictor of the lesson planning competency (Hypothesis 3).

4.2.5.2 Effects of Scaffolding on Lesson Planning Competency

The first research question aimed at estimating the effects of scaffolding focused on facilitating the formulation of learning goals during observational learning on the lesson planning competency, by comparing the lesson planning competency of control and experimental conditions at the post-test. There was no significant difference for the lesson planning competency between the experimental and the control conditions in the post-test and

the scores in goal formulation obtained during the delayed post-test (hypothesis 1 was rejected). Table 4.2 presents means, standard deviations and ANOVA F-tests of the comparisons of experimental and control groups. The repeated measures ANCOVA (including test, condition and their interaction) also did not identify significant gain in lesson planning competency $F(4, 160) = 0.52$, $p = 0.72$.

Table 4.2.
Means, SD and F-test statistics for the pre-, post- and delayed post-tests

	Experimental M (SD)	Control M (SD)	F-test statistic (1,80) (p-value)
Pre-test	-0.008 (0.54)	0.007 (0.64)	0.012 (0.91, n.s.)
Post-test	-0.037 (0.72)	0.141 (0.61)	1.46 (0.23, n.s.)
Delayed post-test	0.219 (0.29)	0.167 (0.21)	0.79 (0.38, n.s.)

There was no statistically significant difference in lesson planning competency between experimental and control conditions detected during pre-test, post-test and the score in formulating learning goals in delayed post-test. Delayed post-test controlled for pre-test scores (ANCOVA) did not show significant differences in formulating learning goals by experimental and control conditions either (F (1,75) = 0.75, $p = 0.39$).

4.2.5.3 Predictors of lesson planning competency at the post-test phase

The second and third research questions were focused on identifying the role of additional factors (teaching experience, beliefs about the importance of learning goals and adherence to instructions during the treatment) in predicting lesson planning competency in pre-service teachers from experimental and control conditions. Neither teaching experience ($p = 0.24$) nor beliefs about the importance of learning goals ($p = 0.10$) were statistically significant predictors of lesson planning competency, therefore hypotheses 2a and 2b were rejected.

The multiple regression model with pre-test lesson planning competency and adherence to the instruction during observation in experimental condition, explained 10.7% of variance (Adjusted R-squared) in post-test lesson planning competency: $F(2, 40) = 3.51$, $p = 0.04$. Adherence to the instructions during the observation was the only statistically significant ($p = 0.02$) positive predictor of post-test lesson planning competency, $B = 0.013$, $SE = 0.005$, $Beta = 0.35$. Pre-test lesson planning competency was a positive, but not statistically significant ($p = 0.25$) predictor, $B = 0.19$, $SE = 0.16$, $Beta = 0.17$ (hypothesis 3 was accepted). The overall multiple regression model in both experimental and control conditions, including pre-test lesson planning competency and adherence to the instruction was not significant, which implies that there were other factors beyond pre-test lesson planning competency and adherence to instructions, responsible for the variation in post-test lesson planning competency.

4.2.6 Conclusions

The empirical study adopted pre-post control group design to assess (1) the effect of scaffolded observational learning on the pre-service teachers´ lesson planning competency; (2) the role of teaching experience and beliefs about the importance of learning goals in prediction the post-test lesson planning competency; (3) the role of adherence to instructions during treatment phase in prediction of the post-test lesson planning competency. To answer the first research question ANCOVA analysis was used to compare the lesson planning competency of the experimental and the control conditions at pre-test and post-test phases. It was hypothesised that the conditions would not differ from each other at the pre-test phase, but the experimental condition would be significantly higher in lesson planning competency at the immediate and delayed post-test. To answer the second and third research question, multiple linear regression analysis was used to identify statistically significant predictors of the post-test lesson planning competency. It was hypothesised that prior teaching experience, beliefs about the importance of learning goal formulation and the adherence to instructions during the treatment phase would predict the lesson planning competency at the post-test phase.

The results indicated that although no statistically significant difference was found between treatment and control groups in the lesson planning competency at the post-test phase, adherence to the instruction was a significant predictor of the post-test lesson planning competency of the experimental group. This implies that learning to formulate learning goals (by scaffolding observational learning from example lessons) had a positive effect on lesson planning competency, but the intervention in its current form was not sufficient to achieve significantly higher lesson planning competency of the experimental condition.

Among the limitations of the empirical study were following (1) short intervention phase, focused on one aspect of lesson planning, probably not enough to notice the differences in the lesson planning competency, (2) relatively low level of responding to the intervention (only few students actually completed the tasks during intervention), different type of instructional support might be needed (3) high variance of lesson planning competency within the control and the experimental conditions, implying another possible factor influencing the lesson planning competency. Further research might focus on ensuring more adherences to the instructions, or even explore, what predicts this adherence to design optimal instructions and scaffolds. The interventions in the following studies might be longer, and focus on multiple elements of the lesson planning competency.

5 General Discussion

This doctoral dissertation was conducted in the domain of Teacher Education and aimed to contribute to understanding of the effectiveness of instructional approaches (specifically observational learning) to foster teaching skills and competencies. The dissertation was built up on Bandura´s social-cognitive learning theory (1986), the learning framework for the use of the observational learning suggested by Chi and colleagues (2008), as well as the cognitive skill acquisition theory proposed by Van Lehn (1996), In regard to defining and measuring the teaching competence, the current thesis consideredthe approaches of Stoof et al. (2002) as well as that of Blömeke et al. (2015). The overarching research questions of the dissertation were to identify (1) if observational learning is an effective teaching/learning strategy that contribute to fostering the pre-service teachers´ competence; and (2) in what way should observational learning be organised to ensure that target competency is acquired in the most effective way. The observational learning was shown to be an instructional method able to foster various teaching skills, but also lesson planning competency in pre-service teachers as more general construct if properly scaffolded.

5.1 Summary of the studies

5.1.1 The Meta-Analysis on the Effects of Observational Learning in Teacher Education

The meta-analysis was conducted in the domain of teacher education and grounded on Bandura´s socio-cognitive theory of learning (1986). Its aim was to systematically review empirical studies focused on using observational learning to acquire teaching related skills (using specific teaching strategies, classroom management techniques, etc.). The analysis focused on learning the complex cognitive skills which were demonstrated during observational learning phases, and the subsequent performance of the demonstrated skill. This went in line with Chi´s (2009) definition of observational learning as the process of acquiring

(from observing the model) and further demonstrating an observed skill, as well as with the model of the cognitive skills acquisition suggested by Van Lehn (1996).

The study investigated empirical findings of the studies in Teacher Education. The research questions raised in the meta-analysis were aimed at (1) identifying the effects of observational learning on acquisition of teaching skills, which were measured on objective and subjective scales; and (2) the role of design features (presentation format, measures of performance) and (3) scaffolding in moderating these effects.

The findings supported the theoretical framework in observational learning and previous research. With a relatively small sample size of 19 empirical studies, the meta-analysis was able to identify significantly high effect of observational learning on acquisition of teaching skills (both measured on objective and subjective scales) by pre-service teachers. The summary effect size from 13 studies reporting comparisons regarding objective measures was g = 1.13, *CI* [0.72, 1.54]. The effects of the objective measures showed high heterogeneity I^2 = 91.97%.The summary effect from 6 studies reporting comparisons regarding subjective measures was g = 1.07, *CI* [0.60, 1.54]. The analysis also showed moderate heterogeneity I^2 = 41.87% in the effects measured by subjective measures. Further moderator analyses were performed only for effects measured by objective measures to clarify what factors had contributed to the effectiveness of observational learning for acquisition of teaching skills. Moreover, on the basis of a comprehenisve range of statistical methods to detect and correct for publication bias and questionable research practices, it could be conluded that the identified effects were not due to the bias, and that observational learning is a powerful technique to facilitate complex skill acquisition in pre-service teachers.

The moderator analysis indicated several factors influencing the effects of observational learning on teaching related learning outcomes. Considering the presentation format of learning material, the use of video- and text-based materials yielded similar results (g = 1.06 for video and g = 1.09 for text). Although this contradicts to some extent the

commonly supported idea of the benefits of using video-based learning materials to facilitate learning, it also shows that both presentation formats have unique strengths and weaknesses and both can be an effective way to present the learning material. Interestingly, the “in vivo observation” presentation format significantly outperformed both video- and text. This might be due to higher levels of involvement in the “in vivo observation” setting, but the evidence is limited to only a few studies, thus the interpretation should be treated with caution. More studies using in-vivo observation to present the learning material are needed.

Concerning measures of performance, assessing performance with written measures provided less heterogeneous effects, compared to using assessment of actual performance; however, both can indicate the learning reliably and with the similar magnitude. This finding appears to be due to a method-effect, as the criteria for written measures can be more structured and require less decision making from the examiners. Actual performance, in contrast, might have many features, not covered in coding manual and require more subjective decisions from examiners On the other hand, actual performance reflects the complexity of the learning/teaching situation, but some other factors, like individual style or personal characteristics (not being part of the assessment scale) might influence the outcomes.

The findings concerning the use of additional instructions and scaffolding show that scaffolding increases learning from observation (Chi et al., 2008; Dianovsky & Wink, 2011; Glogger et al., 2009; Hübner, 2009; Stegmann et al., 2012; Van Gog & Rummel, 2010). It was not possible to determine the best combination of support to facilitate learning due to the relatively few studies in each scaffolding scheme (namely, providing scaffolding during observation, provide instruction after observation, using instructional support continuously and not using any additional instruction). Nevertheless, the findings go in line with the already existing empirical evidence, that the continuous scaffolding (having activities both during and after observation phase) was a more effective way of scaffolding pre-service teachers. In contrast to previous research, observational learning, which was not supported by additional

instructions or scaffolding was also effective for skill acquisition. However, this finding might be due to the fact that not all the researchers conducting empirical studies elaborated on the procedure of implementing observation in their studies and therefore, information about the instructions which were actually given was lacking. Alternatively, this finding can be explained by the fact that target skills could be well observed and no special instruction was needed to learn them. The tendencies shown in the findings should be interpreted with caution. More empirical studies investigating different scaffolding schemes are required for future syntheses.

Interestingly, most of the studies in the meta-analysis were focused on the acquisition of a single teaching strategy or technique, rather than addressed the teaching competency on a higher level. To identify if observational learning is an effective instructional method which can target acquisition of more complex constructs like competence, an empirical study was conducted. It allowed testing if scaffolded observational learning could be effective to foster development of the lesson planning competency as a more complex construct, one which integrates a range of knowledge and skills.

5.1.2 Measuring Lesson Planning Competency: The Scale Development

The scale development was conducted as an intermediate step to prepare for an empirical study in the domain of Teacher Education. The empirical studies included in the meta-analysis focused on the acquisition of single teaching principles or techniques, but not on teaching competence in general. It did not suggest an instrument to measure the teaching competence which considers its complexity, so a new measurement scale was developed.

Prior to the development of the scale, it was decided to develop and validate a written measure of performance, as the preceding meta-analysis had indicated, that the written measures should be as effective as actual performance measures, but at the same time that it can be implemented on the larger group of participants and requires less resources in terms of time and cost. It was also decided to focus on lesson planning competency as a part of teaching competence for two main reasons. Firstly because lesson planning competency, as essential

part of teaching activity across different domains, consists of similar components as the teaching competence in general (pedagogical content knowledge, analytic and planning skills, etc.). Secondly, lesson planning competency was considered to relate more to the tasks that teachers perform rather than to their person related characteristics. This implies that it is a capacity easier to foster within shorter periods of time and also that the possible changes in performance can be attributed to the treatment with a higher degree of confidence. Another decision made before the scale was developed was to focus on cognitive skills and processes (noticing, analysing, planning), considering that motivational factors, as another building block of competency should be measured separately with other scales. The scale was developed in the domain of physical education, due to the fact that teachers´ activity and students´ performance can be directly observed. Furthermore, there is little research in the effective teachers´ training in the domain of physical education and this research can therefore contribute to the respective community. Nevertheless, the principles of scale definition and construction can be used to create similar scales in other domains.

The scale was constructed using Item Response Theory and validated on the data collected from two subsequent samples of pre-service teachers in 2016 and 2017. The data allowed to construct two different, but comparable measurement scales for two video clips (with 48 and 50 items respectively), that can be used to assess the lesson planning competency at two different time points (i.e., pre and post-test). Both scales indicated good fit to the one-parameter Rasch model, which in turn provided evidence that lesson-planning competency can be addressed as a unidimensional construct, even though it requires content and pedagogic knowledge, consists of several skills and underlying processes (i.e., noticing important lesson organisation, analysing classroom situation, matching observed units with theory, formulating learning goals, making decisions about effective teaching strategies and equipment used, planning own lessons).

The constructed scales reached the acceptable level of reliability, i.e. they provided enough information to distinguish between different levels of lesson planning competency. Video clip 1 reached WLE Reliability of 0.67, and EAP Reliability of 0.69, and Video clip 2 reached WLE and EAP Reliability of 0.70. To assess the gain in lesson planning competency, the ability scale for both video clips was standardised for the pre-test by extracting the mean value from each individual ability score.

The detailed description of the scale construction, item selection and the general considerations about measuring the competency contribute to the research in teaching competence and can be adapted to other domains. The main consideration about the scale to be made before applying it to lesson planning competency measurement is that motivational aspects of this competency are not included in the scale and should be controlled separately. The developed scale and its variations adapted to other teaching related competencies can contribute to an understanding of how observational learning can foster teachers´ knowledge and skill acquisition at the level of competency and facilitate the development of Teacher Education, as these scales consider the complexity of the teaching and allow to assess competency level and its change due to applied instructional interventions.

5.1.3 Fostering Lesson Planning Competency in Pre-Service Teachers

The empirical study was conducted in the domain of Teacher Education. This study aimed at addressing the question if observational learning combined with scaffolding would have a positive effect on the lesson planning competency of pre-service teachers in physical education. The study was designed to contribute to empirical studies underrepresented in meta-analysis (i.e. used in-vivo observation, used scaffolding during observation). The study was conducted during the Basic Qualification Sports course with 84 pre-service teachers as participants sample in 2017. The pre-post design with random assignment to experimental (N=43) and control condition (N=41) was applied to identify if the treatment (fostering goal formulation during observational learning) would have a positive effect on the lesson planning

competency of pre-service teachers. The scale developed to measure lesson planning competency (see Chapter 3) was used to capture this competency before and after the treatment. The goals of the study were to (a) assess the effect of scaffolding focused on facilitating the formulation of learning goals during observational learning on the acquisition of lesson planning competency, (b) assess the role of teaching experience and teaching beliefs on the lesson planning competency, and (c) assess the role of adherence to the instructions during the observation and its effect on lesson planning competency.

In contrast to expectations, the results indicated that there was no significant difference in lesson planning competency between the experimental and control condition neither in the immediate nor in the delayed post-test. Teaching experience and the teachers' beliefs about the importance of setting learning goals for effective lesson planning were not significant predictors of lesson planning competency.

As expected, adherence to the instructions during the observation was a statistically significant positive predictor of post-test lesson planning competency; furthermore it was the only significant predictor. A multiple regression model with pre-test lesson planning competency and adherence to the instruction during observation in the experimental condition, explained 10.7% of variance (Adjusted R-squared) in post-test lesson planning competency. Overall, the findings of this empirical study highlight the need of instructional support for observational learning to foster lesson planning competency.

5.2 Integration of findings

In respect to overarching goals and research questions of the doctoral dissertation, the results from meta-analytical and empirical studies support the hypothesis that observational learning is an effective teaching/learning strategy that contributes to fostering the pre-service teachers´ competence. It has a positive effect on acquisition of teaching techniques and single didactic principles, but also has potential to foster lesson planning competency as a more complex construct, which combines different types of knowledge and skills. In respect to

organisation and design of observational learning the following statements can be made. Firstly using videos and texts as examples to observe the target skill or behavior can both lead to effective skill acquisition; in vivo observation is potentially the most beneficial for of presenting the target skill or behavior, but more research is needed to check this assumption. Secondly findings of this dissertation suggest that more scaffolding for pre-service teachers result in better skill acquisition (findings from meta-analysis). Moreover, scaffolding contributes to skill or competence acquisition only if pre-service teachers show adherence to the instructions during treatment phase. To sum up, observational learning as an instructional method appears capable of addressing the challenges and requirements towards developing teaching competence. Empirical findings of the meta-analysis provided evidence that learning from observations (as worked examples, modelling target behavior) has a substantial positive effect on acquisition of complex cognitive skills in general, social interaction skills and might foster a professional competence development if properly scaffolded.

5.3 Limitations of the studies

One of the common limitation for all the studies in this dissertation is that the studies were focused on the pre-service teachers and the findings cannot be generalized to in-service teachers, who have more experienced, and probably to some extend different needs and expectations from the courses and programs in further education. Although, the search for primary studies for the meta-analysis was not limited to pre-service teachers, no eligible studies were found, which would provide evidence for the use of observational learning in in-service teacher education and its effects on their knowledge and skill acquisition. This issue generally reflects the state of empirical research in Teacher Education.

The scale designed in Chapter 3 had better discrimination ability on the lower levels of competency and therefore could fail to discriminate between teachers with initially high level of lesson planning competency. In other words, it is also limited to the use for pre-service

teachers only.. Adding several more items with high difficulty, would significantly increase the reliability of the scale and would also make it appropriate for more experienced teachers.

The main limitation of the empirical study is the relatively low level of responsiveness in the intervention. Only few students actually completed the tasks during intervention, and that could be one of the reasons for the treatment not reaching significant effects. One of other possible reasons is rather short intervention phase, focused on one aspect of the lesson planning. It was probably not enough to capture a change in the lesson planning competency. The high variance of lesson planning competency within the control and the experimental conditions, imply other possible factors influencing the lesson planning competency. The solution for this limitation would be to plan a longer intervention addressing several aspects of lesson planning, and using different types of instructional support which would probably be more accepted by the participants. More information about pre-service teacher beliefs, motivation, and preferred learning styles might explain the variance which was not captured in the empirical study presented in this dissertation. Moreover, a more complex intervention would need a more complex scale (probably, with several dimensions) to assess the gain in lesson planning competency. Such a scale would also allow to identify which component or process of lesson planning competency is the most problematic and needs more instructional support.

5.4 Theoretical implications

This doctoral dissertation contributed to research on observational learning in Teacher Education by conducting a systematic review and meta-analysis on the effects of observational learning as instructional method on acquisition of teaching skills, which were modelled during the observation. The results go in line with the previous research conducted in teacher education (Allen & Ryan, 1969; Darling-Hammond et al., 2005; Santagata, Zannoni, & Stigler, 2007) and other domains (Couzijn, 1999; Groenendijk et al, 2013; Rummel & Spada, 2005; Schworm & Renkl, 2007; Van Steendam et al, 2010); and provide evidence for the high effects

of this instructional method for acquisition of teaching skills. The meta-analysis also investigated the role of design-related features (presentation format of the target skill, use of scaffolding and additional instructional support, use of different measures to assess learning outcomes), which has practical and theoretical implications as it contributes to the knowledge about the impact of scaffolding on learning. The findings go in line with studies on scaffolding (Chi et al., 2008; Dianovsky & Wink, 2011; Van Gog & Rummel, 2010) and support the idea that at the lower level of prior knowledge (i.e. for pre-service teachers) more scaffolding result in better skill acquisition. However, to identify the best amount and combination of instructional support aids and scaffolding more empirical research is needed.

The meta-analytic study also contributes to the methodological discussion about publication bias and questionable research practices by discussing strengths and weaknesses of existing methods, namely Egger's test (Sterne & Egger, 2001), Trim'n'fill (Duval & Tweedie, 2000), p-curve analysis (Simonsohn, Nelson, & Simmons, 2014), R-index (Schimmack, 2012), and the fail-safe N (Rosenthal, 1979). Their use in combination to address replicability and generalizability of the research in Teacher Education should inspire their implementation in other domains to obtain evidential value of the effects.

The scale developed within this doctoral dissertation contributed to the understanding of the structure of lesson planning competency as one of the essential parts of teaching competence. The recent research (Blömeke et al, 2015, Seidel & Stürmer, 2014) suggests several dimensions to be considered in measuring competence. However, the study in scale development demonstrates that a lesson planning competency can be considered a unidimensional construct, although it involves several underlying processes and skills.

The empirical study contributes to the research in the field of physical education, which has not been sufficiently researched so far. It also contributes to the body of empirical research by using the design features underrepresented in the sample of studies included in the meta-

analysis and therefore can be used in future systematic reviews and meta-analyses in the domain of Teacher Education.

5.5 Further research

One of the directions for further development is performing a research synthesis to identify the factors (related to the design of learning activities, learning outcomes, experience and motivation of the participants, different teaching domains) that make observational learning effective instructional method to foster teaching competence. Although, observational learning is an effective method to foster acquisition of single teaching strategy or principles, it also has the potential to be beneficial for achieving more complex learning goals and tasks if properly supported. This direction requires well-designed empirical studies that use observational learning and scaffolding to foster teaching skills and competencies. More research should be done using in-vivo observation, but also other presentation formats, to determine the most effective way to foster teaching competence. According to the findings of the systematic review and the meta-analysis, different schemes of scaffolding (providing instructional support at different stages of learning or continuously) are also underrepresented in current empirical research; knowing the strengths and weaknesses of different types and amount of scaffolding would be beneficial for designing educational programs.

In general, focus on transfer of observational learning experiences to later application of observational learning by pre- and in-service teachers is the promising direction of further research. To support this direction of the research creating the reliable, objective scales to measure the construct of teaching competence as a whole becomes a central issue. The item response theory approach seems to be a suitable method, as it allows to consider the difficulty of the items and better reflect the structure of competence as a complex construct. Developing methodologically similar scales for different disciplines and domains would enable identifying what kind of competencies can be best fostered with observational learning and which might need other instructional approaches. Creating the pool of measures to assess teaching

competence and its components in different domains would also contribute a lot in designing and conducting empirical studies, but also could help to unify measures used in the Teacher Education research. Therefore, it would provide an opportunity to attribute differences between the effects of observational learning (or other methods) to differences in treatment and design, but not to measurement errors. And in turn would promote a better connection between theoretical research and its implementation.

Referring to the evidence from the empirical study on fostering lesson planning competence in pre-service teachers and difficulties encountered in this study, there are a few more ideas and considerations to be implemented in future research. For example, making sure that students follow the instructions during treatment should be one of the main issues. Researchers could treat adherence to the instructions as an ability and search for predictors to optimize the instruction for further studies.

5.6 Practical implications

The studies conducted within this doctoral dissertation provide evidence for observational learning to be a powerful instructional method, which can be used in teacher education to promote a variety of teaching skills and competencies (acquisition of specific didactic principles, teaching strategies, lesson planning competency, etc.). Observational learning has a positive effect on learning if outcomes are measured by objective scales (knowledge tests, performance), but is also highly acceptable by students (subjective measures about perceived learning and motivation to apply acquired skills) which has not only theoretical, but also the practical significance for further research and practice in the domain of Teacher Education.

The first implication considers the design of empirical studies and learning environments implementing observational learning as an instructional method for pre-service teachers to acquire teaching skills and competencies. The observational learning procedures can be integrated into education programs and be suggested for bigger groups of pre-service

teachers, without losing in the authenticity of the context and quality of the skills acquired. For example, according to the findings of the meta-analysis both video and text presentation can be used with similar effectiveness in skill acquisition. In vivo observations have even more potential to establish authentic and effective learning. Different schemas for providing instructional support and scaffolding, presented in the systematic review and meta-analysis chapter open up different options for designing the learning programs, with the consideration that observational learning might work out even with minimal scaffolding. Furthermore, more scaffolding has higher effects on skill acquisition.

The second possible practical implication is related to assessing the teaching competence and therefore the effectiveness of programs or seminars directed at acquisition of teaching skills. The chapter on the scale development provides insights about structure of teaching competence and underlying competencies and should encourage researchers and practitioners to create assessment instruments that comprise the complexity and interconnection of teaching skills. The methodological sections provide a step-by-step guidance to define and establish the scale, together with an example of such scale using the Item Response Theory. The discussion and the justifications for decision making about items in the scale can be used to create similar scales for other skills and competencies.

In closing, the evidence presented and discussed in this dissertation, supports the ideas of learning by observation, suggested by Albert Bandura (1986) over 30 years ago and develops them to meet challenges in fostering teaching competence nowadays. The combination of theoretical and empirical research on professional competence, scaffolding, use of technology in education, and also advances in research methodology and scale development techniques contribute significantly to providing insights for further research and practical implications of observational learning as instructional method.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31* (2-3), 162-172. DOI: 10.1016/j.stueduc.2005.05.008

Allen, D., & Ryan, K. (1969). *Microteaching.* Reading, MA: Addison-Wesley.

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Balz, E. (2008). Fachdidaktische Konzepte oder: Woran soll sich der Schulsports orientieren. In E. Balz & P. Wolters (Eds.), *Schulsports. Didaktik und Methodik,* (pp. 34-40). Leipzig: Erhard Friedrich Verlag GmbH.

Bandura, A. (1986). *Social Foundations of Thought and Action*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive theory*. Upper Saddle River, NJ: Prentice-Hall.

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual Review Psychology, 52* (1), 1-26. DOI: 10.1146/annurev.psych.52.1.1

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of educational research, 74* (1), 29-58. DOI: 10.3102/00346543074001029

Bayerische Staatsregierung (2008). *Ordnung der Ersten Prüfung für ein Lehramt an öffentlichen Schulen. Lehramtsprüfungsordnung I - LPO I.* Fassung vom 13.08.2008. Online verfügbar unter: http://www.gesetze-bayern.de/Content/Document/BayLPO_I-119, retrieved in June 2017.

Berthold, K., Nuckles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction, 17* (5), 564-577. DOI: 10.1016/j.learninstruc.2007.09.007

Beswick, K., Muir, T. (2013). Making connections: Lessons on the use of video in pre-service teacher education. *Mathematics Teacher Education and Development, 15*(2), 27–29. Retrieved from: https://eric.ed.gov/?id=EJ1018707

Blömeke, S., Gustafsson, J.E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie, 223* (1), 3□13. DOI: 10.1027/2151-2604/a000194

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Borko, H., Shavelson, R. J., Stern, P. (1981). Teachers' Decisions in the Planning of Reading Instruction. *Research Quarterly, 16* (3), 449-466. DOI: 10.2307/747411

Braaksma, M. A. H., Rijlaarsdam, G., & Van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology, 94* (1), 405–415. DOI: 10.1037/0022-0663.94.2.405

Carroll, J.B. (1993). *Human cognitive abilities.* A survey of factor-analytic studies. New York: Cambridge University Press.

Castellan, C.M. (2010). Quantitative and qualitative research: a view for clarity, *International Journal of Education, 2* (2), 1-14. Retrieved from: www.macrothink.org.ije.

Chi, M.T. & VanLehn, K. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47* (3), 177-188. DOI: 10.1080/00461520.2012.695709

Chi, M.T. (2009). Active-Constructive-Interactive: A conceptual framework of differentiating learning activities. *Topics in Cognitive Science, 1* (1), 73-105. DOI: 10.1111/j.1756-8765.2008.01005.x

Chi, M.T., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32* (2), 301-341. DOI: 10.1080/03640210701863396

Chomsky, N. (1968). *Language and Mind*. New York: Harcourt, Brace & World, Inc..

Choppin, B. H. (1997). Objective tests. In J. Keeves (Ed.), *Educational research, methodology and measurement: an international handbook* (2nd ed. pp. 354-358): Elsevier Science Ltd.

Cisero, C. A. (2006). Does reflective journal writing improve course performance? *College Teaching, 54* (2), 231-236. DOI: 10.3200/CTCH.54.2.231-236

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction, 9* (2), 109–142. DOI: 10.1016/S0959-4752(98)00040-1

Cree, V., & Macaulay, C. (Ed.) (2000). *Transfer of Learning in Professional and Vocational Education*. Routledge.

Crissman, J. K. (2006). The design and utilization of effective worked examples: A meta – analysis. (Unpublished doctoral thesis). ETD collection for University of Nebraska - Lincoln. Retrieved from: http://digitalcommons.unl.edu/dissertations/AAI3208114

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16* (1), 297-334. DOI: 10.1007/BF02310555.

D'Agostino, J. V., Powers, S. J. (2009). Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis. *American Educational Research Journal, 46* (1), 146 –182. DOI: 10.3102/0002831208323280

Darling-Hammond, L., Hammerness, K., Grossman, P., Rust, F., & Shulman, L. (2005). The design of teacher education programs. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 390-441). San Francisco: Jossey-Bass.

Dianovsky, M. T., & Wink, D. J. (2012). Student learning through journal writing in a general education chemistry course for pre-elementary education majors. *Science Education, 96* (3), 543-565. DOI: 10.1002/sce.21010

Duval, S., Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56* (2), 455-463. DOI: 10.1111/j.0006-341X.2000.00455.x

Epstein, R. M., Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of American Medical Association 287* (2), 226–235. DOI: 10.1001/jama.287.2.226.

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist, 28* (1), 401-416.

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist, 48* (1), 56-66. DOI: 10.1080/00461520.2012.748005

Froschmeier, T., Feimeier, T., Friedl, Y., Schneiderbanger, M., Spitzenpfeil, B., Weiss, G. (2016). *Moderner Sportsunterricht in Stundenbildern 3/4*. Augsburg: Auer Verlag.

Froschmeier, T., Stegmann, K., Zottmann, J., & Matikalo-Siegl, K. (2012). *Instructional support for vicarious learning in teacher education*. Paper presented at the EARLI.D4. Present & Discuss Session: Learning environments.

Fryling, M. J., Johnson, C., Hayes, L., (2011). Understanding Observational Learning: An Interbehavioral Approach. *The Analysis of Verbal Behavior, 27* (1), 191-20 PMCID: PMC3139552 3.

Gaudin, C., & Chalies, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review 16* (1), 41-67. DOI: 10.1016/j.edurev.2015.06.001

Glogger, I., Holzapfel, L., Schwonke, R., Nückles, M., Renkl, A., (2009). Activation of Learning Strategies in Writing Learning Journals: The Specificity of Prompts Matters. *Zeitschrift fuer Paedagogische Psychologie, 23* (2), 95–104. DOI: 10.1024/1010-0652.23.2.95

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & Van den Bergh, H. (2013). The effect of observational learning on students' performance, processes, and motivation in two creative domains. *British Journal of Educational Psychology, 83* (1), 3-28, DOI:10.1111/j.2044-8279.2011.02052.x

Grossman, P. L. (1990). *The Making of a Teacher. Teacher Knowledge and Teacher Education.* New York: Columbia University, Teachers College Press.

Hanson, J., Bannister, S., Clark, A., & Raszka, W. (2010). Oh, What You Can See: The Role of Observation in Medical Student Education. *Pediatrics, 126* (5), 843 -845. DOI: 10.1542/peds.2010-2538

Henninger, J. C. (2002). The Effects of Knowledge of Instructional Goals on Observations of Teaching and Learning. *Journal of Research in Music Education, 50* (1), 37-50. DOI: 10.2307/3345691

Herling, R. W. (2000). Operational definitions of expertise and competence. *Advances in Developing Human Resources, 2* (1), 8-21. DOI: 10.1177/152342230000200103

Hiebert, J. & Stigler, J. W. (2017) Teaching Versus Teachers as a Level for Change: Comparing a Japanese and a U.S. Perspective on Improving Instruction. *Educational Researcher, 46* (4), 169–176. DOI: 10.3102/0013189X17711899

Heitzmann, N., Fischer, F., Kühne-Eversmann, L., & Fischer, M. R. (2015). Enhancing diagnostic competence with self-explanation prompts and adaptable feedback. *Medical education, 49*(10), 993-1003. DOI: 10.1111/medu.12778

Higgins, J. P., Thompson, S. G., Deeks, J. J., Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327* (7414), 557–560. DOI: 10.1136/bmj.327.7414.557

Hoover, D. J., Giambatista, R. C., & Belkin, L. J. (2012). Eyes On, Hands On: Vicarious Observational Learning as an Enhancement of Direct Experience. *Academy of Management Learning & Education*, *11*(4), 591-608. DOI: 10.5465/amle.2010.0102

Hübner, S., (2009). Learning journals as medium of self-regulated learning: How to design instructional support to overcome strategy deficits? (Unpublished doctoral dissertation). Albert-Ludwig University, Freiburg. Retrieved from: http://www.freidok.uni-freiburg.de/volltexte/7179/pdf/Diss_RP_huebner_finale_UB.pdf

Huitt, W. (2004). Observational (social) learning: An overview. [Online] *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved on 25.02.2015 from http://www.edpsycinteractive.org/topics/soccoq/soclm.html

Ingvarson, L., & Rowe, K. (2008). Conceptualising and evaluating teacher quality: substantive and methodological issues. *Australian Journal of Education*, *52*(1), 5-35. DOI: 10.1177/000494410805200102

Jensen, L. (2001). Planning lessons. In M. Celce-Murcia (ed.) *Teaching English as a Second or Foreign Language*, (pp. 403-408). Boston, MA: Heinle & Heinle.

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modelling and assessment. *Zeitschrift für Psychologie / Journal of Psychology*, *216* (2), 60–72. DOI: 10.1027/0044-3409.216.2.61

Krischner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based learning. *Educational Psychologist, 41*(2), 75-86. DOI: 10.1207/s15326985ep4102_1

Lai, E. (2011). Motivation: A literature review research report. Retrieved July 18, 2017 from http://images.pearsonassessments.com/images/tmrs/Motivation_Review_final.pdf

Lau, J., Ioannidis, J., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*(1), 597 – 600. DOI:10.1136/bmj.333.7568.597

Lefstein, A., Snell, J. (2011). Professional vision and the politics of teacher learning. *Teaching and Teacher Education, 27* (3), 505-514. DOI: 10.1016/j.tate.2010.10.004

Marini, A., & Genereux, R. (1995). The Challenge of Teaching for Transfer. In A. McKeough, J. Lupart, & A. Marini (Eds.), *Teaching for Transfer: Fostering Generalization in Learning* (pp. 1-20). New Jersey: Lawrence Erlbaum Associates.

McClelland, D.C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28* (1), 1-14.

Menekse, M., Stump, G.S., Krause, S., & Chi, M.T.H. (2013). Differentiated overt learning activities for effective instruction in an engineering classroom. *Journal of Engineering Education, 102* (3), 346–374. DOI: 10.1002/jee.20021

Miller, G.E. (1990). The assessment of Clinical Skills/Competence/Performance. *Academic Medicine, 65* (9), 63-67.

Orwin, R.G. (1983) A Fail-Safe N for Effect Size in Meta-Analysis. *Journal of Educational Statistics, 8* (2), 157-159. DOI: 10.2307/1164923.

Patel, V. L., Kaufman, D.R. & Magder, S.A. (1996). The acquisition of medical expertise in complex dynamic environments. In K.A. Ericsson (Ed.), *The road to excellence*. (pp. 127-165) Mahwah, NJ: Erlbaum.

Perkins, D. N., Salomon, G. (1992). Transfer of Learning. Contribution to the International Encyclopedia of Education, Second Edition. Oxford, England: Pergamon Press.

Raedts, M., Rijlaarsdam, G., Van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs,

task knowledge and writing performances. In G. Rijlaarsdam (Series Ed.), & P. Boscolo, & S. Hidi (Vol Eds.), *Studies in writing, 19* (1), 219–238. Oxford: Elsevier

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction, 16* (2), 92-103. DOI: 10.1016/j.learninstruc.2006.02.008

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: a cognitive load perspective. *Educational Psychologist, 38* (1), 15-22. DOI: 10.1207/S15326985EP3801_3

R-Index.org (2014). R-Index 1.0. www.r-Index.org.

Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal of Vocational Training, 40* (1), 123–139. Retrieved from: https://files.eric.ed.gov/fulltext/EJ776614.pdf

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86* (3), 638-641. DOI: 10.1037/0033-2909.86.3.638

Rothstein, J. M. (1989). On defining subjective and objective measurements. *Physical Therapy, 69* (7), 577-579. DOI: 10.1.1.1014.2022

Rummel, N. & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *Journal of the Learning Sciences, 14* (2), 201–241. DOI: 10.1207/s15327809jls1402_2

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 Math Teacher Education, 10* (2), 123–140. DOI: 10.1007/s10857-007-9029-9

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17* (4), 551-566. DOI: 10.1037/a0029487

Schimmack, U. (2016). The replicability-index: Quantifying statistical research integrity. Retrieved from https://wordpress.com/post/replication-index.wordpress.com/920

Schwonke, R., Wittwer, J., Alven, V., Salden, R., Krieg, C., & Renkl, A. (2007). *Can tutored problem solving benefit from faded worked-out examples?* Paper presented at the The European Cognitive Science Conference, Delphi, Greece. DOI: 10.1.1.68.3711

Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99* (2), 285– 296. DOI: 10.1037/0022-0663.99.2.285

Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal, 51* (4), 739-771. DOI: 10.3102/0002831214531321

Shavelson, R. J, Stern, P. (1981). Research on Teachers Pedagogical Thoughts, Judgements, Decisions, and Behavior. *Review of Educational Research, 51* (4), 455-498. DOI: 10.3102/00346543051004455

Shavelson, R. J. (2010). On the measurement of competency. *Empirical research in vocational education and training, 2* (1), 41□63. Retrieved from: www.pedocs.de urn:nbn:de:0111-opus-52350.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420-428. DOI: 10.1037/0033-2909.86.2.420

Shulman, L. S. (1987). Knowledge and Teaching. Foundations of the New Reform. *Harvard Educational Review, 57* (1), 1-23. DOI: 10.17763/haer.57.1.j463w79r56455411

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-Curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143* (2), 534-547. DOI: 10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-Curves: Making P-Curve Analysis More Robust to Errors, Fraud and Ambitious P-Hacking, A Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144* (6), 1146-1152. DOI: 10.1037/xge0000104

Sonnenschein, S., & Whitehurst, G. J. (1984). Developing referential communication: A hierarchy of skills. *Child Development, 55* (5), 1936–1945. DOI: 10.2307/1129940

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method. *Journal of Organizational Behavior, 15* (5), 385-392. DOI: 10.1002/job.4030150503

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and instruction*, *21*(1), 22-33. Retrieved from https://www.learntechlib.org/p/108435.

Stegmann, K. (2015). MyMetaAnalysis. Programm zur Durchführung von Metaanalysen. Retrieved from: http://www.karsten-stegmann.de

Stegmann, K., Pilz, F., Siebeck, M., & Fischer, F. (2012). Vicarious learning during simulations: Is it more effective than hands-on training? *Medical Education, 46* (10), 1001-1008. DOI: 10.1111/j.1365-2923.2012.04344.x

Sterne, J. A., Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology, 54* (10), 1046-1055. DOI:10.1016/S0895-4356(01)00377-8

Stoof, A., Martens R. L., van Merriënboer J. G, .and Bastiaens, T. J. (2002). The Boundary Approach of Competence: A Constructivist Aid for Understanding and Using the Concept of Competence. *Human Resource Development Review 1 (*3), 345-365. DOI: 10.1177/1534484302013005

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19-30). NY: Cambridge University Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). The redundancy effect. In J. M. Spector & S. P. Lajoie (Eds.) *Cognitive Load Theory*, (pp. 141 – 154). Springer NY. DOI:10.1007/978-1-4419-8126-4

Tanner-Smith, E., Tipton, E., Polanin, J. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: a Tutorial in R. *Journal of Developmental and Life-Course Criminology, 2* (1), 85-112. DOI: 10.1007/s40865-016-0026-5

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–184). Lawrence Erlbaum Associates Publishers.

Thornbury, S. (1990). *Metaphors we work by: EFL and its metaphors.* Paper presented at the ALAA/ ALS Conference, Macquarie University, Sydney, Australia. Retrieved from http://203.72.145.166/ELT/files/45-3-1.pdf

Van Der Vleuten, C. P. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education: Theory and Practice, 1* (1), 41-67. doi: 10.1007/BF00596229

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22* (2), 155–174. DOI: 10.1007/s10648-010-9134-7

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20* (4), 316–327. DOI: 10.1016/j.learninstruc.2009.08.009

VanLehn, K. (1996). Cognitive Skill Acquisition. *Annual Review of Psychology, 47 (1),* 513-539. DOI: 10.1146/annurev.psych.47.1.513

Verhelst, N. D. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online, 6 (3),* 231–271. Retrieved from: https://www.dgps.de/fachgruppen/methoden/mpr-online/issue13/issue15/art2/verhelst.pdf

Wang, X. (2013). A Potential Approach to Support Pre-service Teachers' Professional Learning: The Video Analysis of the Authentic Classroom. *US-China Education Review 3*(3), 149-161.

Weber, H. & Westmeyer, H. (1998). *Die Inflation der Intelligenzen*. Vortrag gehalten auf dem 41. Kongreß der Deutschen Gesellschaft für Psychologie: Dresden, 28.9.-1.10.1998.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L.H. Salganik (Eds.), *Defining and Selecting Key Competencies*, (pp. 45-65). Göttingen: Hogrefe & Huber.

Zottmann, J.M., Stegmann, K., Strijbos, J-W., Vogel, F., Wecker, C., Fischer, F. (2013). Computer-supported collaborative learning with digital video cases in teacher education: The impact of teaching experience on knowledge convergence. *Computers in Human Behavior, 29* (5), 2100–2108. DOI:10.1016/j.chb.2013.04.014

Primary studies for the meta-analysis*

*Bloch, M. N. (1977). Incidental Observation of Multiple Teaching Models as a Form of Training. (Unpublished doctoral thesis) Stanford University, USA. Retrieved from: University Microfilms International, Ann Arbor, Michigan, 48106, USA.

*Claus, K.E. (1969). *Effects of Modeling and Feedback Treatments on the Development of Teachers' Questioning Skills*. Technical Report No. 6, Stanford Center for Research and Development in Teaching. Retrieved from: ERIC Document Reproduction Service (Document Number: ED 033081).

*Crooks, C., Gifford, V. D. (1992). A Comparison of Videotaped Teaching Models and the Lecture Technique in Increasing the use of Questioning Strategies Presented in Elementary Science Lessons. *Journal of Science Teacher Education, 3*(3), 76-78. Retrieved from http://www.jstor.org/stable/43155943

*Gettinger, M., Stoiber, K. C. (2014). Increasing opportunities to respond to print during storybook reading: Effects of evocative print-referencing techniques. *Early Childhood Research Quarterly, 29* (3), 283–297. Retrieved from: https://doi.org/10.1016/j.ecresq.2014.03.001

*Haverback, H. R., Parault, S. J. (2011). High efficacy and the preservice reading teacher: A comparative study. *Teaching and Teacher Education, 27* (4), 703–711. DOI: 10.1016/j.tate.2010.12.001

*Koran, J. J., Jr. (1969). The Relative Effects of Classroom Instruction and Subsequent Observational Learning on the Acquisition of Questioning Behavior by Pre-Service Elementary Science Teachers. *Journal of Research in Science Teaching, 6* (3), 217-223. DOI: 10.1002/tea.3660060305

*Koran, J. J., Jr. (1970). A Comparison of the Effects of Observational Learning and Self-Rating on the Acquisition and Retention of Questioning Behavior by Elementary

Science Teacher Trainees. *Science Education, 54* (4), 385-389. DOI: 10.1002/sce.3730540414

*Koran, J. J., Jr., Koran M. L., McDonald, F. J. (1972). Effects of Different Sources of Positive and Negative Information on Observational Learning of a Teaching Skill. *Journal of Educational Psychology, 63* (5), 405-410. Retrieved from: http://dx.doi.org/10.1037/h0033242

*Koran, M. L., Snow, R. E. and McDonald, F. J. (1971). Teacher Aptitude and Observational Learning of a Teaching Skill. *Journal of Educational Psychology, 62* (3), 219-228. Retrieved from: http://dx.doi.org/10.1037/h0031142

*Kubany, E. S., Sloggett, B. B. (1991). Attentional factors in observational learning: Effects on acquisition of behavior management skills. *Behavior Therapy, 22* (3), 435-448. Retrieved from: https://doi.org/10.1016/S0005-7894(05)80376-4

*Lavin, K. T. (1992). Testing the Effects of Observational Learning Theory on Teaching Behavior and Collegiality. (Unpublished doctoral thesis) Fordham University, New York. Retrieved from: University Microfilms International, Ann Arbor, Michigan, 48106, USA.

*Lee, Y., Ertmer, P. A. (2006). Examining the Effect of Small Group Discussions and Question Prompts on Vicarious Learning Outcomes. *Journal of Research on Technology in Education, 39* (1), 66-80. DOI: 10.1080/15391523.2006.10782473

*Moreno, R., Valdez, A. (2007). Immediate and Delayed Effects of Using a Classroom Case Exemplar in Teacher Education: The Role of Presentation Format. *Journal of Educational Psychology, 99* (1), 194-206. DOI: 10.1037/0022-0663.99.1.194

*Slogget, B. B. (1972). The Comparative Effects of Verbal Information, Passive Observation, and Active Observation on the Acquisition of Classroom Management Skills. (Unpublished doctoral dissertation). University of Hawaii. Retrieved from: *ProQuest Dissertations & Theses Global.*

*Wang, L., Ertmer, P. A. (2003). Impact of Vicarious Learning Experiences and Goal Setting on Preservice Teachers' Self-Efficacy for Technology Integration: A Pilot Study. Paper presented at the annual meeting of the *American Educational Research Association*, Chicago, April 21-25, 2003.

R-packages used in the analyses

Fox, J. (2017). Package "car". Retrieved on 01.12.2017 from https://cran.r-project.org/web/packages/car/car.pdf

Irribarra, D. T., Freund, R. (2016). Package "WrightMap". Retrieved on 12.03.2017 from https://cran.r-project.org/web/packages/WrightMap/WrightMap.pdf

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T. (2016). Package "eRm". Retrieved on 01.12.2016 from https://cran.r-project.org/web/packages/eRm/eRm.pdf

Pinheiro, J. (2017). Package "nlme". Retrieved on 12.03.2017 from https://cran.r-project.org/web/packages/nlme/nlme.pdf

R Core Team (2017). Package "stats". Retrieved on 01.10.2017 from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

Rizopoulos, D. (2017). Package "ltm". Retrieved on 01.10.2017 from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Robitzsch, A., Kiefer, T., Wu, M. (2017). Package "TAM". Retrieved on 01.12.2017 from https://cran.r-project.org/web/packages/TAM/TAM.pdf

Viechtbauer, W. (2014). Package "metafor". Retrieved on 02.10.2015 from: http://cran.r-project.org/web/packages/metafor/metafor.pdf

Wickham, H. (2017). Package "stringr". Retrieved on 12.03.2017 from https://cran.r-project.org/web/packages/stringr/stringr.pdf

Willse, J. T. (2017). Package "CTT". Retrieved on 01.12.2017 from https://cran.r-project.org/web/packages/CTT/CTT.pdf

APPENDIX

Informed consent for the data collection 2017

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR PSYCHOLOGIE UND PÄDAGOGIK
DEPARTMENT PSYCHOLOGIE
LEHRSTUHL FÜR EMPIRISCHE PÄDAGOGIK UND PÄDAGOGISCHE PSYCHOLOGIE

Einverständniserklärung

Sie erhalten ein Exemplar dieser Einverständniserklärung für Ihre Unterlagen.

Teil 1. Information über die Verwendung der erhobenen Daten
Die Studie beschäftigt sich mit Lernen durch Beobachten und zielt darauf ab, Instruktionsmethoden herauszuarbeiten, die optimale Unterstützung für Referendare zu Verfügung stellen. Die erhobenen Daten dienen einzig Forschungszwecken um zukünftige Kurse „Basisqualifikation Sport" zu verbessern.

Vertraulichkeit
Die erhobenen Daten werden vertraulich behandelt, weiterverarbeitet und haben keinerlei Einfluss auf die Bewertung von Studienleistungen. Um eine größtmögliche Anonymisierung zu gewährleisten, bitten wir Ihnen einen Code statt Ihr Vorname und Nachname zu benutzen.

Verweigerung oder Widerruf des Einverständnisses
Ihnen entstehen **keinerlei** Nachteile im Rahmen ihrer Studiums, wenn Sie Ihr Einverständnis verweigern. Sie können Ihr Einverständnis außerdem jederzeit widerrufen. Der Widerruf ist elektronisch per e-mail zu richten an Olga Chernikova (e-mail: O.Chernikova@campus.lmu.de). Bitte der unten gegebene Code als Referenz benutzen.

Teil 2. Schriftliche Einverständniserklärung
Die von mir im Rahmen der Befragung am 3. Apri 2017 erhobenen Daten dürfen für Forschungszwecke verwendet werden.

BITTE NUR ankreuzen wenn Sie NICHT einverstanden sind.

☐ Ich willige **nicht** ein, dass die erhobenen Daten für Forschungszwecke verwendet werden dürfen.

Ich habe die Informationen in Teil 1 gelesen und verstanden. Ich hatte die Gelegenheit, Fragen zu stellen, und meine Fragen wurden in diesem Fall zu meiner Zufriedenheit beantwortet. Meine Einwilligung in die Speicherung der erhobenen Daten zu den oben angegebenen Zwecken geschieht freiwillig.

Code: Erste Buchstabe des Namens Ihrer Mutter ___
Erste Buchstabe des Namens Ihres Vaters ___
Ihr Geburtsdatum (z.B. 07 bzw. 17) ___

Datum: | 0 | 3 | . | 0 | 4 | . | 2 | 0 | 1 | 7 |

Unterschrift: _____

BQS course overview (in German)

Basisqualifikation Sports - Grundschule 03.04. bis 07.04.2016

Zeit	Montag	Dienstag	Mittwoch	Donnerstag	Freitag
9.15 - 10.00	***9.00 Uhr*** *„Check in“*- wichtige Informationen zur Lehrgangswoche HAW	*„Radmethodik“* TH IRL	Koordinative Fähigkeiten DÜRR	Unterrichtsmitschau FÜR	*„Auf und nieder“* Kräftigungszirkel für Kids RAT
10.00 - 10.45	*„Die erste Sportstunde“* Spielerisches Bewegen und Orientieren in der Sporthalle DÜR/HAW	*„Handstand“* TH IRL	Noch offen DÜRR	*“Bumball”* Innovative Spielformen mit Bällen HAW	*„Laufen so schnell und wendig wie ein Gepard“* Kreative Laufspiele LaH RAT
10.45 - 11.00	**PAUSE**	**PAUSE**	**PAUSE**	**PAUSE**	**PAUSE**
11.00 - 11.45	*Pädagogik und Didaktik des Sportunterrichts an Grundschulen* FRO	*„Meine Sprungwelt“* Sicherheitstipps und Unterrichtsideen mit dem Minitrampolin TH IRL	*“Über Stock und über Stein”* TH “Balancieren, Klettern, Stützen an Großgeräten“ WUN	*„Gigaball“* Innovative Spielformen mit Bällen HAW	*„Ganz schön weit geflogen“* Methodik zum Weitspringen LaH RAT
11.45 - 12.30	*„Wild spielen“* Grundlagen der Ballspielentwicklung FRO	IRL	*“Abenteuersafari”* TH "Kräftigung der Arm- und Rumpfmuskulatur an Gerätebahnen" WUN	*„Lang aber langsam!“* Ausdauerndes Laufen RAT	*„Power Hour“* Bewegung nonstop im Rhythmusparcours FRO
12.30 - 13.30	**PAUSE**	**PAUSE**	**PAUSE**	**PAUSE**	**PAUSE**
13.30 - 14.15	Unterrichtsmitschau Körperkontaktspiele SCH	*“Das kann ich schon”* Lernfortschritte mit dem Kann-Buch begleiten MAE	*„Völkerball- und Brennballspiele“* Spannende und aktivierende Varianten BIL	*“Mens sana in corpore sano“* Bewegungsangebote für den Schulalltag RAT	*„Durchs wilde Kurdistan“* Sinnvolle Staffelspiele FRO
14.15 -15.00	Unterrichtsmitschau Basketball Schaustunde SCH	*Strategiespiele* „Geier und Takeshi“ MAE	*„Robinson Crusoe“* Tennisähnliche Rückschlagspiele FRO	*„Bin ich im Gleichgewicht?“* Bewegungsaufgaben mit u. o. Partner RAT	
15.00 - 15.15		**PAUSE**	**PAUSE**	**PAUSE**	**PAUSE**
15.15 - 16.00	*Die Spiel-in-Echt-Methode* Spielerisches Technik- und Taktiklernen FRO	*“Vertrau in deine Kräfte”* faires Ringen und Kämpfen MAE	*„My Style“* Coole Moves für Kids STR	Unterrichtsorganisation für Handballspiele FÜR/HAW	
16.00 - 16.45	*Die Spiel-in-Echt-Methode* Spielerisches Technik- und Taktiklernen FRO	*“Schlag den Ball nicht mich!”* Spiele mit Schlägern MAE	*“Werfen x Fangen!”* Jonglage und mehr STR	Tschoukhandball FÜR	

Background information: data collection 2017

Demograpische Daten

Geschlecht: ☐ männlich ☐ weiblich

Alter: _____ Jahre

Fachsemester: _____

Angestrebtes Lehramt: ☐ Grundschule ☐ Förderschule

Bereits absolvierter Umfang des Erziehungswissenschaftlichen Studiums (EWS): ca. _____% (Anzahl Kurse: ____)

Haben Sie Erfahrungen im Geben von Sportsunterricht oder als Trainer/Trainerin (o.ä.): ☐ Nein ☐ Ja

Wenn Ja, wie haben Sie diese Erfahrung gesammelt?
(Was haben Sie gemacht? Wie lange haben Sie diese Tätigkeit ausgeübt? Haben Sie dafür eine spezielle Ausbildung absolviert?):

_____

_____

_____

Stellen Sie sich vor, Sie sollen eine Unterrichtsstunde zum Thema Weit- und Hochsprung *planen*. Welche der folgenden Aussagen halten Sie für besonders wichtig bzw. weniger wichtig für die *Planung* der Unterrichtsstunde? Vergeben Sie Zahlen zwischen 1 (am wichtigsten) und 5 (am wenigsten wichtig bzw. unwichtig):

_____ Am Ende der Stunde springen sie SuS 10% weiter als in der letzten Stunde.

_____ Die SuS üben die korrekte Sprungtechnik.

_____ Die Lehrkraft demonstriert die korrekte Sprungtechnik.

_____ Die SuS erhalten Informationen zu verschiedene Sprungtechniken.

_____ Die SuS wärmen sich vor dem Sprungtraining durch 5 Runden Laufen auf.

Pre- and post-test questionnaires (in German)

Code: Erste Buchstabe des Vornamens Ihrer Mutter___
Erste Buchstabe des Vornamens Ihres Vaters ___
Ihr Geburtsdatum (z.B. 07 bzw. 17) ___

Bandfarbe _____

Aufgabe

Sie werden ein 3-minütiges Video sehen, welches einen Ausschnitt aus einer Sportsstunde zeigt. Danach werden Sie 12 Minuten Zeit haben, um einige Fragen zur Videosequenz zu beantworten - die empfohlene Antwortzeit ist bei jeder Frage angegeben. Bevor die Videosequenz beginnt, haben Sie einige Minuten Zeit, sich die Fragen durchzulesen.

Bitte geben Sie eine kritische Bewertung für die Stundensequenzen ab, die Sie gesehen haben. Bitte beziehen Sie Ihre Kommentare dabei auf konkrete Momente in dem Video (z.B. "als die Lehrkraft zeigt, wie…") und erklären Sie Ihre Gedanken kurz.

a. Bitte geben Sie bis zu drei Lernziele an, die die Lehrkraft für die Sequenz gesetzt haben könnte. (2 Min.)

_____

_____

_____

_____

_____

b. Bitte geben Sie bis zu 3 von der Lehrkraft verwendete Lernstrategien an, die Sie als besonders effektiv betrachten und bis zu 3 solche, die Ihnen als schwach vorkamen. Bitte kurz begründen. (3 Min.)

_____

_____

_____

_____

_____

_____

_____

_____

c. Wie würden Sie den Unterricht fortführen (Nächste Lernschritte, Lernsequenzen)? Denken Sie an mögliche Übungen die das geplante Lernziel unterstützen können. (2 Min.)

_____

_____

_____

_____

Bitte beantworten Sie die folgenden 10 Fragen zur Videosequenz (5 Min.)

1. Listen Sie die bitte die bei der/n Übung/en verwendeten Materialien / Geräte auf:

_____

_____

_____

Bitte kreuzen Sie die passende/n Antworte/n an (Mehrfachnennungen möglich).

2. Welche Stundensequenz(en) haben Sie in dem Video gesehen?

- ☐ Einstimmung
- ☐ Hauptteil
- ☐ Aufwärmen
- ☐ Ausklang

3. Welches Ziel verfolgte(n) die Übung(en), die Sie in dem Video gesehen haben?

- ☐ Krafttraining
- ☐ Koordinationstraining
- ☐ Ausdauertraining
- ☐ Sozialkompetenzen

4. Gebrauch der Sportshalle (SH)

- ☐ nur ein kleiner Teil der SH wird benutzt
- ☐ (fast) die ganze SH wird von einer Gesamtgruppe benutzt
- ☐ Teile der SH werden von mehreren Untergruppen simultan benutzt
- ☐ Teile der SH werden von mehreren abwechselnd für unterschiedliche Übungen benutzt

5. Bewegungsgelegenheiten

- ☐ Die Schüler bewegen sich ständig/stoppen die Bewegung nur um neue Aufgaben zu erhalten
- ☐ Aktive Bewegungsphasen wechseln mit ruhigen Zuhörphasen
- ☐ Es gibt längere Zeitabschnitte, während welcher einige Schüler nicht aktiv sind
- ☐ Die Schüler haben wenig Bewegungsgelegenheiten

6. Welche Lernmethoden sind für die Phase benutzt?

- ☐ Analytisch-synthetische Methode (vom Speziellen zum Ganzen)
- ☐ Ganzheitsmethode (es wird an der ganzen Aufgabe gearbeitet)
- ☐ Die induktive Methode (Exploration durch Schüler)
- ☐ Die deduktive Methode (genaue Anleitung Schritt bei Schritt)

7. Wie wurde für Sicherheit gesorgt?

- ☐ Sicherheitsvorkehrungen waren nicht erkennbar
- ☐ Spezialausrüstung wurde benutzt
- ☐ Sicherheitsanweisungen /-regeln wurden im Voraus gegeben
- ☐ Die Lehrkraft zeigte die Bewegungsabläufe im Voraus
- ☐ Die Lehrkraft gab Sicherheitsanweisungen während der Übung

8. Komplexität der Übungen

- ☐ Zu einfach für die Schüler
- ☐ Entsprach den physischen Voraussetzungen und dem Alter der Schüler
- ☐ Zu komplex für die Schüler
- ☐ Die Schüler konnte die Komplexitässtufe je nach ihren Voraussetzungen zu wählen
- ☐ Die Komplexität steigerte sich im Verlauf der Übung(en)

9. Kreativität der Schüler

- ☐ Die Kinder haben Möglichkeit die Bewegungen selbst auszudenken
- ☐ Alle Bewegungen wurden vorgezeigt
- ☐ Die Kinder stellen eigene Ideen (Übung, Tanz, Station) vor
- ☐ Die Kinder sind engagiert eigene Erfahrung in der Besprechung zu teilen

10. Reaktionen der Schüler.

- ☐ Die Lehrkraft bat die Schüler um Feedback
- ☐ Negative Emotionen wurden gleich beachtet
- ☐ Die Schüler sahen glücklich und zufrieden aus
- ☐ Die Schüler sahen perplex aus
- ☐ Die Schüler waren aktiv und gerne bei der Übung dabei
- ☐ Die Schüler sahen gelangweilt aus

Vielen Dank für Ihre Teilnahme!

Observation forms: Introduction for the control condition (in German)

Code: Erste Buchstabe des Vornamens Ihrer Mutter____
Erste Buchstabe des Vornamens Ihres Vaters ____
Ihr Geburtsdatum (z.B. 07 bzw. 17) ____

Beobachtungsbogen zur Unterrichtsmitschau

Sie werden gleich die Gelegenheit haben eine Doppelstunde einer Grundschulklasse zu beobachten. Die Doppelstunde ist in zwei Themen unterteilt: (1) Körperkontaktspiele und (2) Basketball. Zu jeder der beiden Teile finden Sie in diesem Heft je zwei Seiten. Auf der letzten Seite haben Sie darüber hinaus die Möglichkeit, weitere allgemeine Notizen zu machen. Die Beobachtung der Schulstunde soll Ihnen helfen in Zukunft eigene Sportsstunden zu planen und durchzuführen. Bitte denken Sie daher daran, beim Anfertigen der Beobachtungsnotizen auf Aspekte zu achten, die Ihnen später bei ihrer eigenen Planung und Durchführung helfen könnten. Fokussieren Sie dabei hier zunächst nur die Aspekte die für die Planung der konkreten Schulstunde relevant sind.

Bundesarchiv, B 145 Bild-F010151-0007
Foto: Steiner, Egon | 29. April 1961

Am Ende der Unterrichtsmitschau werden wir dieses Heft wieder einsammeln. Sie erhalten das Heft am Freitag wieder zurück. Bitte denken Sie daran den Code oben auf dieser Seite auszufüllen, da wir nur dann in Lage sind, Ihnen Ihren Beobachtungsbogen zurückzugeben. Vielen Dank für Ihre Unterstützung!

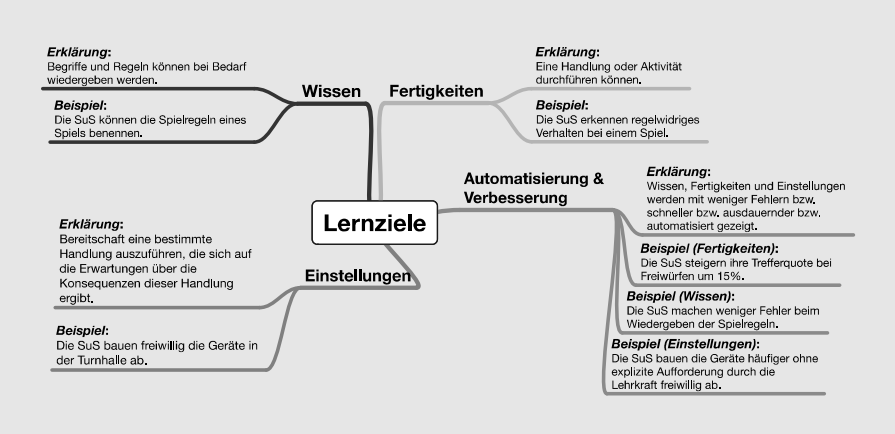
Observation forms: Introduction for the experimental condition (in German)

Code: Erste Buchstabe des Vornamens Ihrer Mutter____
Erste Buchstabe des Vornamens Ihres Vaters ____
Ihr Geburtsdatum (z.B. 07 bzw. 17) ____

Beobachtungsbogen zur Unterrichtsmitschau

Sie werden gleich die Gelegenheit haben eine Doppelstunde einer Grundschulklasse zu beobachten. Die Doppelstunde ist in zwei Themen unterteilt: (1) Körperkontaktspiele und (2) Basketball. Zu jeder der beiden Teile finden Sie in diesem Heft je zwei Seiten. Auf der letzten Seite haben Sie darüber hinaus die Möglichkeit, weitere allgemeine Notizen zu machen. Die Beobachtung der Schulstunde soll Ihnen helfen in Zukunft eigene Sportsstunden zu planen und durchzuführen. Bitte denken Sie daher daran, beim Anfertigen der Beobachtungsnotizen auf Aspekte zu achten, die Ihnen später bei ihrer eigenen Planung und Durchführung helfen könnten. Fokussieren Sie dabei hier zunächst nur die Aspekte die für die Planung der konkreten Schulstunde relevant sind.

Nach den Vorgaben des Lehrplans soll Unterricht vom Ziel her geplant werden. Daher ist es notwendig, dass Sie sich Gedanken darüber machen welche Lernziele im beobachteten Unterricht erreicht werden sollen. Dazu müssen Sie sich die Frage stellen „Was sollen die Schüler am Ende des Unterrichts (besser) können oder wissen als zu Beginn der Stunde?“. Das Lernziel beschreibt also ein Ergebnis (z. B. „Die SuS können Abseitsstellungen beim Fußball erkennen.“) und nicht eine Aktivität der Schüler oder der Lehrkraft während des Unterrichts (z. B. „Die SuS übernehmen im Unterricht abwechselnd die Rolle des Schiedsrichters.“, „Die Lehrkraft erklärt die Abseitsstellung.“)

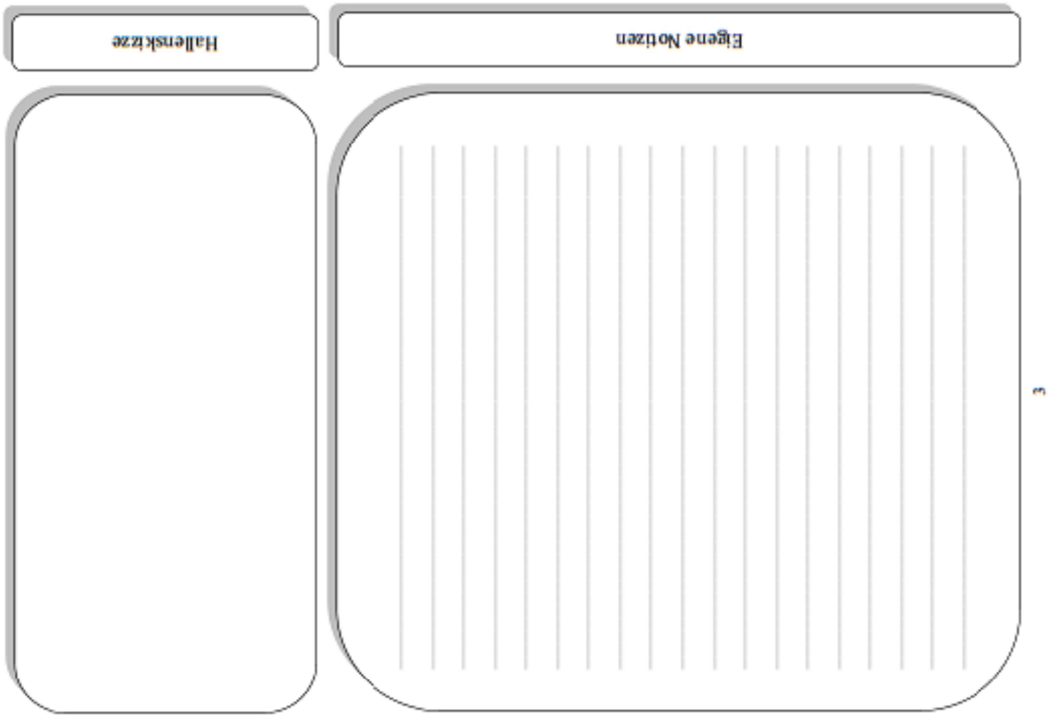


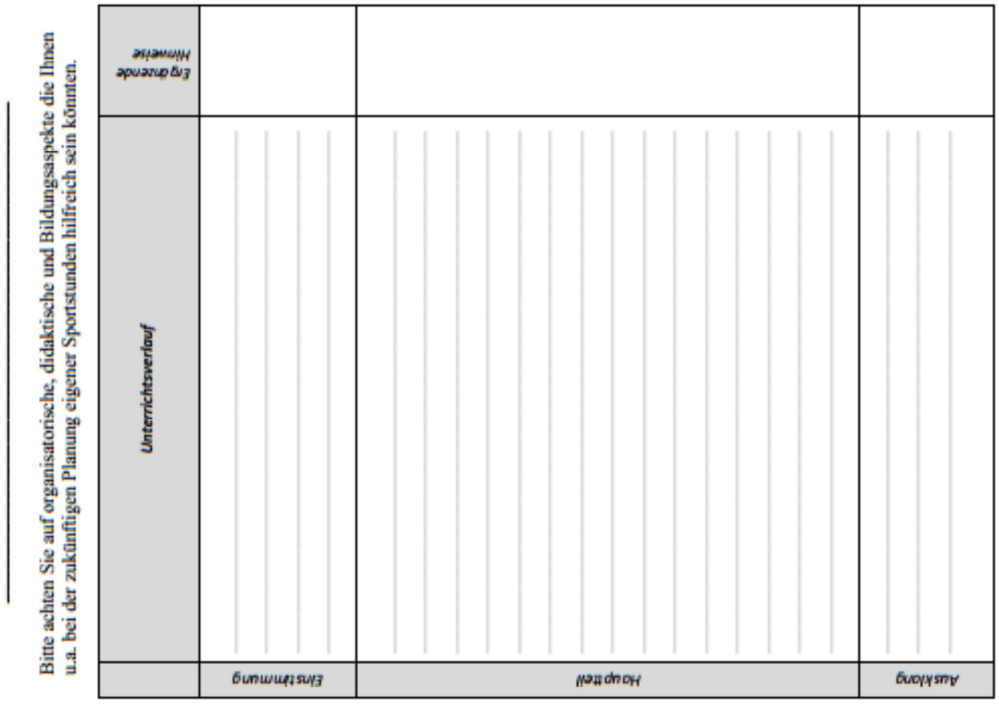
Ein gutes Lernziel muss an bestimmtem Verhalten der Schüler erkennbar sein. Daher sollten zu allgemeine Lernziele (z. B. „Die Schülerinnen und Schüler verstehen das Konzept der Fairness.“) vermieden werden. Stattdessen sollte das Lernziel beobachtbar bzw. messbar formuliert werden (z. B. „Die Schülerinnen und Schüler können Abseitsstellungen erkennen.“). Das Lernziel spezifiziert dabei möglichst konkret die Aktivität/Fertigkeit die Schülerinnen und Schüler durchführen können bzw. zeigen sollen.

Am Ende der Unterrichtsmitschau werden wir dieses Heft wieder einsammeln. Sie erhalten das Heft am Freitag wieder zurück. Bitte denken Sie daran den Code oben auf dieser Seite auszufüllen, da wir nur dann in Lage sind, Ihnen Ihren Beobachtungsbogen zurückzugeben. Vielen Dank für Ihre Unterstützung!

Example of Learning Diary Page

(The same design was also used in observation forms and delayed post-test)





Observation forms: scaffolding task for the experimental condition (in German)

Inserted instead of “Eigene Notitzen”, see Appendix X

Hauptlernziel der Schulstunde

Was können die Schülerinnen und Schüler im Anschluss an die Schulstunde (besser als zu Beginn der Unterrichtsstunde)?

Woran könnte man beobachten, dass das Lernziel erreicht wurde?
Wie könnte man messen, inwieweit das Lernziel erreicht wurde?

Bitte nummerieren Sie in Ihren Notizen zum Unterrichtsverlauf (linkes Blatt) die einzelnen Unterrichtsaktivitäten und beantworten Sie dazu folgende Frage: ***Inwieweit diese Aktivitäten zum Erreichen des Lernziels beigetragen haben*?** Vergeben Sie dazu Bewertungen von
-3 (diese Aktivität war kontraproduktiv) bis
+3 (diese Aktivität unterstützte das Erreichen des Lernziels). Die Wertung 0 sollten Sie vergeben, wenn Sie denken, dass eine Aktivität nicht zum Erreichen des Lernziels beiträgt, aber auch keinen negativen Effekt hat.

No.	Bewertung	No.	Bewertung	No.	Bewertung
___	_____	___	_____	___	_____
___	_____	___	_____	___	_____
___	_____	___	_____	___	_____
___	_____	___	_____	___	_____
___	_____	___	_____	___	_____

Planungshilfe

Delayed post-test: introduction for both conditions (in German)

Code: Erste Buchstabe des Vornamens Ihrer Mutter____
Erste Buchstabe des Vornamens Ihres Vaters ____
Ihr Geburtsdatum (z.B. 07 bzw. 17) ____

Beobachtungsbogen zur Unterrichtsmitschau

Sie werden gleich die Gelegenheit haben eine Sportsstunde einer Grundschulklasse zu beobachten. Sie finden in diesem Heft zwei Seiten für Notizen zu dieser Sportsstunde. Auf der letzten Seite haben Sie darüber hinaus die Möglichkeit, weitere allgemeine Notizen zu machen. Die Beobachtung der Schulstunde soll Ihnen helfen in Zukunft eigene Sportsstunden zu planen und durchzuführen. Bitte denken Sie daher daran, beim Anfertigen der Beobachtungsnotizen auf Aspekte zu achten, die Ihnen später bei ihrer eigenen Planung und Durchführung helfen könnten. Fokussieren Sie dabei hier zunächst nur die Aspekte die für die Planung der konkreten Schulstunde relevant sind.

Am Ende der Unterrichtsmitschau werden wir dieses Heft wieder einsammeln. Sie erhalten das Heft am Freitag wieder zurück. Bitte denken Sie daran den Code oben auf dieser Seite auszufüllen, da wir nur dann in Lage sind, Ihnen Ihren Beobachtungsbogen zurückzugeben. Vielen Dank für Ihre Unterstützung!

Eidesstattliche Versicherung
Statement of Scientific Integrity

Chernikova Olga
Name, Vorname
Last name, first name

Ich versichere, dass ich die an der Fakultät für Psychologie und Pädagogik der Ludwig-Maximilians-Universität München zur Dissertation eingereichte Arbeit mit dem Titel:
I assert that the thesis I submitted to the Faculty of Psychology and Pedagogy of the Ludwig-Maximilian-Universität München under the title:

What Makes Observational Learning in Teacher Education Effective?

Evidence from a meta-analysis and an experimental study

selbst verfasst, alle Teile eigenständig formuliert und keine fremden Textteile übernommen habe, die nicht als solche gekennzeichnet sind. Kein Abschnitt der Doktorarbeit wurde von einer anderen Person formuliert, und bei der Abfassung wurden keine anderen als die in der Abhandlung aufgeführten Hilfsmittel benutzt.
is written by myself, I have formulated all parts independently and I have not taken any texts components of others without indicating them. No formulation has been made by someone else and I have not used any sources other than indicated in the thesis.

Ich erkläre, das ich habe an keiner anderen Stelle einen Antrag auf Zulassung zur Promotion gestellt oder bereits einen Doktortitel auf der Grundlage des vorgelegten Studienabschlusses erworben und mich auch nicht einer Doktorprüfung erfolglos unterzogen.
I assert I have not applied anywhere else for a doctoral degree nor have I obtained a doctor title on the basis of my present studies or failed a doctoral examination.

München, 02.05.2018

Ort, Datum
Place, Date

Olga Chernikova

Unterschrift Doktorandin/Doktorand
Signature of the doctoral candidate