

**Meta-analysis on the impact of in-service professional development programs for
preschool teachers on quality ratings and child outcomes**

Inaugural Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Franziska Egert

aus

Dachau

Bamberg, den 30.01.2015.

Tag der mündlichen Prüfung: 06.07.2015

Dekan: Universitätsprofessor Dr. Stefan Hörmann

Erstgutachter: Universitätsprofessor Dr. Hans-Günther Roßbach

Zweitgutachter: Universitätsprofessor Dr. Ruben G. Fukkink

ABSTRACT

In the last decade, several efforts have been made to improve teaching and intervening practice in preschool classrooms by means of professional development programs (Martinez-Beck & Zaslow, 2006; OECD, 2012). In-service professional development for preschool teachers varies widely with respect to amount, duration, delivery format (traditional workshop and course vs. modern approaches such as coaching, mentoring, blended learning) and content (Oberhuemer, 2012; 2013). In general, it is assumed that in-service training enhances teachers' knowledge, skills, and classroom practice. Improved teaching practice enhances student achievement. The connection seems intuitive, but empirical demonstration is difficult. Therefore, the purpose of this investigation was to conduct a research synthesis on the results of in-service professional development effects and to draw conclusions about the findings on potential effect modifiers into an integrated framework.

Systematic review procedures and meta-analytic techniques were used to estimate the effectiveness of in-service programs. The review consists of experimental studies (RCT, CRT, quasi-experimental and non-experimental studies) that have evaluated the impact of professional development for teachers working in center-based care on quality ratings and child development. Electronic searches (ERIC, PsycINFO, ProQuest, FIS, and WISO) and hand searches in renowned journals from 1970 to 2011 produced 1,162 hits. Two independent coders extracted statistical information from the full texts and screened the literature on the quality of the studies, in-service characteristics, and training transfer factors. A randomized multi-level procedure was used to aggregate the summary impact, where effect sizes were nested under treatment.

Overall, 36 studies provided sufficient data to conduct a meta-analysis of in-service training effects on *quality ratings*. The aggregation of findings revealed a medium summary effect $g' = 0.64$ ($SE = 0.071$). Alongside methodological variables, the size of the effect was statistically moderated by training intensity and format. In-service programs with 45 to 60 hours of training and programs that used coaching as their sole format were the most effective.

The impact of professional development programs on the *language and literacy abilities* of young children have been investigated in 48 studies. The summary effect was small $g' = 0.38$ ($SE = 0.045$). Differential impacts and variations in the size of the effects were found for different outcome domains. Effect sizes were moderated by methodological aspects, training design and intensity. Programs with 15 to 30 hours of training were the most effective in fostering language and literacy skills. Approaches that used mixed delivery types, in particular the combination of workshop and course had a larger impact.

For the meta-analysis of in-service training effects on *social skills and reduced problem behavior*, 13 studies provided sufficient data. A small summary effect was found: $g' = 0.45$ ($SE = 0.074$). However, differential effects emerged for different outcome domains. In particular, larger effects were found for the promotion of social skills than for the reduction of problem behavior. Alongside methodological variables, training programs with a duration of 9 months or short-term programs with less than 30 hours of training were significantly more effective.

To investigate the link between in-service professional development, quality improvements and child development, 9 studies were found that provide data on both quality ratings and child development. In-service training effects on student achievement could be significantly predicted by quality improvements that occurred during training. More than 53 % of the variance in effects on child development was explained by the quality ratings which emerged as a result of in-service professional training.

In addition to the meta-analyses, narrative reviews identified some features of highly effective in-service programs with a calculated effect of $g > 0.80$. Narrative analysis indicated that a narrow content focus and educational principles matter. In particular, effective programs provided opportunities for active learning, guidance and feedback as well as performance-based assignments during training and between training sessions. Further, the trainers of highly effective programs were either the authors of the program or had a specific pretraining/license for the program.

Several effect modifiers were tested. In the majority of cases, the effects were influenced by methodological issues (e.g., pretest differences, reliability of scales or informants) and training intensity in hours. However, the results for training delivery variables were inconsistent. Clear conclusions on other characteristics cannot be made from the meta-regression, because the findings were either inconsistent or missing information in primary studies hindered the investigation.

In summary, there are some beneficial programs and the synthesis of research findings suggests that most of the in-service training has positive effects on the quality of early education and child development. The results demonstrate that quality improvements are a key mechanism to enhance student achievement. Some conclusions for achieving highly professional development with regards to duration and intensity, educational principles and trainers can be drawn. Recommendations for further research include high quality research designs (e.g., equivalence of groups, reliable scales, etc.) and data collection on confounding and contextual variables. Nevertheless, some questions remain in analyzing in-service training effects. In particular, more research and primary studies are needed to investigate the heterogeneity of the results and the moderating role of implementation processes and program fidelity.

This thesis is dedicated to my dearly loved grandmother, Anna Egert.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge some of the people who have helped and supported me throughout this incredible journey.

Foremost, I would like to thank my doctoral supervisors Prof. Dr. Hans-Günther Roßbach and Prof. Dr. Ruben G. Fukkink for their time, support and encouragement throughout my doctoral studies. I am extremely grateful for the guidance you have provided me throughout the long and winding process and the rewarding feedback which helped to transform the big idea into an achievable research practice.

The incredible journey started at the FPG Child Development Institute in North Carolina in 2009 while I was conducting my diploma thesis on the effects of professional development with support and guidance from Prof. Dr. Virginia Buysse and Dr. Ellen Peisner-Feinberg. This research visit aroused my interest in professional development research. Three years later, in 2012, I had the chance to discuss my preliminary findings with Prof. Dr. Virginia Buysse, Dr. Pam Winton, and Dr. Ellen Peisner-Feinberg. I appreciate your insightful comments about professional development that reframed my thinking. Working with all of you has been a privilege and a pleasure.

Another step in this incredible journey was Montréal. I would like to express my gratitude to Prof. Dr. Christa Japel for the opportunity to visit Québec. Thank you for your hospitality and encouragement. You have been such a great source of motivation for me. I also would like to thank her husband Robert and her beautiful children Myriam and Lea †, who made Canada my second home.

Further, I want to express my gratitude to Prof. Dr. Andrea G. Eckhardt, who inspired me to conduct systematic reviews and meta-analyses. I appreciate your coding and your willingness to proofread my dissertation, your valuable comments and suggestions, your time and kindness.

I also want to thank and especially recognize my second coder Daniel Turani, whose talent and commitment made this work possible.

I owe my utmost appreciation to my parents, Rosmarie and Werner Egert. Thank you for your boundless love, support, and encouragement. I appreciate your confidence in my abilities to master this challenging journey. You gave me amazing energy to keep moving forward. You have always been there for me, even when I was more than 6,000 miles away.

Last but not least, I would like to gratefully acknowledge Regina Egert for being the most incredible sister and the best bank business manager in the world. Thank you for managing my international banking and insurance business during my journeys.

TABLE OF CONTENTS

1. Introduction	1
1.1. Research objectives	4
1.2. Research procedure	7
2. Theoretical framework and review of relevant literature	10
2.1. Theoretical frameworks.....	10
2.1.1. Process frameworks for professional development effects.....	10
2.1.2. Training transfer factors.....	14
2.1.3. Professional development components and potential effect modifiers	17
2.1.4. Methodological quality and measurement features moderating in-service training effects	27
2.1.5. Integrative framework on professional development effects.....	29
2.2. Review of in-service training effects on quality ratings	35
2.2.1. In-service training impacts on process quality ratings with Environment Rating Scales	41
2.2.2. In-service training impacts on language and literacy-specific quality	44
2.2.3. In-service training impacts on interaction quality ratings with the CLASS	48
2.2.4. In-service training impacts on alternative classroom quality ratings.....	51
2.2.5. In-service training impacts on target-child quality.....	52
2.2.6. Challenges measuring in-service training effects on quality ratings.....	52
2.3. Review of in-service training effects on language and literacy development.....	55
2.3.1. In-service training impacts on language development.....	57
2.3.2. In-service training impacts on literacy development.....	60
2.3.3. Differential effects of in-service training for Dual Language Learners and children at risk	64
2.3.4. Evidence base on in-service training effects on language and literacy.....	67
2.3.5. Challenges measuring in-service training effects on language and literacy.....	71
2.4. Review of in-service training effects on social skills and reduced problem behavior	73
2.4.1. In-service training impacts on social abilities	76
2.4.2. In-service training impacts on reduced problem behavior	78
2.4.3. Challenges measuring in-service training effects on social skills and problem behavior.....	79
3. Methodology	81
3.1. Research aims and problem formulation	81
3.1.1. Research questions.....	81
3.1.2. Selection criteria	83
3.2. Literature Search	84
3.2.1. Electronic search	85
3.2.2. Hand search	85

3.2.3.	Search Process.....	86
3.3.	Coding and gathering information from studies	88
3.3.1.	Coding at study and treatment level	89
3.3.2.	Coding at effect size level.....	94
3.4.	Data extraction and computing effect size	96
3.5.	Aggregation of findings and moderator analysis.....	100
4.	Results.....	104
4.1.	Results on the impact on quality ratings.....	104
4.1.1.	Description of primary studies on quality ratings	104
4.1.2.	Meta-analysis and moderator analysis on quality ratings.....	108
4.1.3.	Narrative review of effective studies on quality ratings.....	114
4.2.	Results on the impact of language and literacy training	117
4.2.1.	Description of primary studies on language and literacy training.....	117
4.2.2.	Meta-analysis and moderator analysis on language and literacy	121
4.2.3.	Narrative review of effective studies on language and literacy.....	128
4.3.	Results on the impact on social-emotional and behavioral skills	130
4.3.1.	Description of primary studies on social-emotional and behavioral skills	130
4.3.2.	Meta-analysis and moderator analysis on social-emotional and behavioral skills.....	133
4.3.3.	Narrative review of effective studies on social-emotional and behavioral skills.....	137
4.4.	Link between in-service training, quality ratings and child development.....	138
4.4.1.	Description of primary studies on quality ratings and child development	139
4.4.2.	Meta-analyses on in-service training effects on quality and child development ...	139
4.4.3.	Association between quality improvements and in-service training effects on child development.....	142
4.5.	Evaluation of in-service training components summarized in the integrative framework.....	143
4.5.1.	Significant training effect modifiers of professional development	144
4.5.2.	Potential effect modifiers from the narrative review	147
4.5.3.	Inconclusive effect modifiers and lack of evidence	148
5.	Discussion.....	150
5.1.	Discussion of results.....	150
5.2.	Limitations	155
5.3.	Implications.....	160
5.3.1.	Professional development research	160
5.3.2.	Professional development for early education teachers.....	162

LIST OF TABLES

Table 1. Frameworks on the process from professional development to child outcomes	11
Table 2. Evidence base on training transfer factors to change workplace performance.....	15
Table 3. Review of studies measuring in-service training impacts on quality ratings	38
Table 4. Evidence base of in-service effects on CLASS ratings.....	49
Table 5. Review of studies measuring in-service training effects on language and literacy	56
Table 6. Evidence base of in-service training effects on language and literacy development	68
Table 7. Review of studies measuring in-service effects on social skill and problem behavior	75
Table 8. Research questions according to outcome domain.....	82
Table 9. Keywords for electronic search	85
Table 10. Literature search	87
Table 11. Inter-rater reliability of study and treatment level	91
Table 12. Inter-rater reliability of potential effect modifiers	93
Table 13. Inter-rater disagreement in percent in coding on effect size level	94
Table 14. Systematization of child outcomes for coding	95
Table 15. Formulas for effect size transformation	99
Table 16. Overview of studies included in meta-analysis on quality ratings	106
Table 17. Meta-analysis on quality ratings	109
Table 18. Moderator analysis of in-service training effects on quality ratings.....	112
Table 19. Overview of studies included in meta-analysis on language and literacy	119
Table 20. Meta-analysis on language and literacy development	122
Table 21. Moderator analysis of in-service training effects on language and literacy domains	124
Table 22. Moderator analysis of in-service effects on language and literacy.....	126
Table 23. Narrative review of effective in-service programs on language and literacy.....	129
Table 24. Overview of studies included in meta-analysis on social and behavioral outcomes	132
Table 25. Meta-analysis on social skills and problem behavior	134
Table 26. Moderator analysis on social and behavioral outcome domains.....	135
Table 27. Moderator analysis of in-service effects on social and behavioral outcomes.....	136
Table 28. Overview of studies on the impact on quality ratings and child outcomes.....	140
Table 29. In-service training impacts on quality ratings.....	141
Table 30. In-service training impacts on child development	141
Table 31. Regression analysis on quality improvements predicting child development.....	142
Table 32. Summary of findings on potential effect modifiers	146

LIST OF FIGURES

Figure 1. Five-step research procedure.....	8
Figure 2. Model for teacher training effects	12
Figure 3. Conceptual framework for professional development	18
Figure 4. Integrative framework on in-service training effects.....	33
Figure 5. Literature search process.....	88
Figure 6. Effect size distribution of quality ratings.....	105
Figure 7. Effect sizes distribution for language and literacy outcomes.....	118
Figure 8. Effect size distribution on social-emotional outcomes	131
Figure 9. Effect size distribution of problem behavior outcomes	131
Figure 10. Evidence-based conclusions on in-service training effect modifiers	144

LIST OF ABBREVIATIONS

ACES	Academic Competence Evaluation Scale
BASC	Behavior Assessment System for Children
BPI	Behavior Problem Index
CAP	Concepts About Print
CBS	Challenging Behavior Scale
CCDC	Creative Curriculum Development Continuum
CCOF	Classroom Competence Observation Form
CG	Control group
CLASS	Classroom Assessment Scoring System
CMA	Comprehensive Meta-analysis Software
CRT	Cluster Randomized Trial
CST	Challenging Situation Task
CTB	CTB Developing Skill Checklist
CTOPPP	Comprehensive Test of Phonological and Print Processing
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
DLL	Dual Language Learner
DSC	Developing Skills Checklist
EAS	Emerging Academic Snapshot
ECBI	Eyberg Child Behavior Inventory
ECCOM	Early Childhood Classroom Measure
ECCOS	Early Childhood Classroom Observation Scale
ECERS	Early Childhood Environment Rating Scale
ECERS-R	Early Childhood Environment Rating Scale Revised Edition
ECERS-E	Early Childhood Environment Rating Scale Extended Version
ECI	Early Communication Indicator
EG	Experimental group
ELAN	Elternfragebogen zur Erfassung des Wortschatzes
ELL	English Language Learner
ELLCO	Early Language and Literacy Classroom Observation
ELS	Early Learning Scale
EOWPT	Expressive One Word Picture Vocabulary Test
ERS	Environmental Rating Scales
FACES	Family and Child Experiences Survey
HS	Head Start
HSET	Heidelberger Sprachentwicklungstest
HUR	HeadsUp! Reading
ICC	Intraclass Correlation Coefficient
ICPS	I Can Problem Solve Curriculum
IDEA	IDEA Oral Language Proficiency Test
IGDI	Individual Growth and Development Indicator
inCLASS	Individualized Classroom Assessment Scoring System
ITERS	Infant Toddler Environment Rating Scale
ITERS-R	Infant Toddler Environment Rating Scale Revised Edition
LELA	Language and Emerging Literacy Assessment
MCDI	MacArthur-Bates Communicative Development Inventory

MLwiN	Multilevel Modeling Software
ORCE	Observational Record of the Caregiving Environment
OWL	Opening the World of Learning
PALS	Phonological Awareness Literacy Screening
PAT	Phonological Awareness Test
PATH	Promoting Alternative Thinking Strategies Curriculum
PBS	Positive Behavior Support
PIPS	Performance Indicator in Primary School
PKBS	Preschool Kindergarten Behavior Scale
PLBS	Preschool Learning Behavior Scale
PLS	Preschool Language Scale
PPVT	Peabody Picture Vocabulary Test
QED	Quasi-experimental design
RCT	Randomized controlled trial
RIGLS	Restricted Maximum Likelihood Variant
SCP	Social Competence Performance Checklist
SDQ	Strength and Difficulties Questionnaire
SELA	Support for Early Literacy Assessment
SELDAK	Sprachentwicklung und Literacy bei deutschsprachig aufwachsenden Kindern
SESBI	Sutter-Eyberg Student Behavior Inventory
SETK	Sprachentwicklungstest für Kinder
SISMIK	Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kitas
SSRS	Social Skills Rating System
TERA	Test of Early Reading Ability
TOLD	Test of Language Development
TOOLS	Tools of Mind Curriculum
TRSSA	Teacher Rating Scale of School Adjustment
TVIP	Test de Vocabulario en Imagenes Peabody (Peabody picture vocabulary test)
WJ	Woodcock Johnson
WLPB	Woodcock Language Proficiency Battery

“How does teacher professional development affect student achievement?

The connection seems intuitive.

But demonstrating it is difficult.”

(Yoon, Duncan, Lee, Scarloss, & Shapley, 2007, p. III)

1. Introduction

There is increased public interest in the professionalization of the early childhood teaching workforce¹ given the augmented policy emphasis on the early years as the foundation for later school success (Buysse, Winton, & Rous, 2009; Sheridan, Edwards, Marvin, & Knoche, 2009). Huge public investments in pre- and in-service professional development for early childhood educators are being made all over the world to improve child care quality and to foster the development of young children in early child care (Martinez-Beck & Zaslow, 2006; Oberhuemer, 2012; 2013; OECD, 2012; Whitebook & Ryan, 2011).

Underlying these investments, there is a consensus that the quality of early childhood education classrooms contributes substantially to children's learning and development (cf. Eurydice, 2014). High-quality center-based care is consistently linked to child development, well-being and the school adjustment of young children in the preschool years (Howes, 1988; McCartney, 1984; National Institute of Child Health and Human Development Early Child Care Research Network [NICHD], 2003; NICHD & Duncan, 2003; Peisner-Feinberg & Burchinal, 1997; Roßbach, 2005; Vandell & Wolfe, 2000). Longitudinal studies also confirm the impact of high-quality child care experiences on later school success (cf. Barnett, 1995; Dickinson & Neuman, 2006; Neuman & Dickinson, 2001; Roßbach, Kluczniok, & Isenmann, 2008; Sammons, 2010; Sammons et al., 2002). Both children with or without disadvantaged family backgrounds profit from high-quality child care (NICHD, 2000a, Peisner-Feinberg & Burchinal, 1997; Sammons, 2010). Experimental studies provide a consistent evidence base for beneficial outcomes of high-quality early childhood programs for disadvantaged children over the course of life (Schweinhart, Barnes, & Weikart, 1993; Schweinhart, 2004; Schweinhart et al., 2005). Relying on those findings, state and federal governments use high-quality early education as a means to address potential achievement deficits in vulnerable children (Bundesministerium für Familien, Senioren, Frauen und Jugend [BMFSFJ], 2012; Burchinal, Hyson, & Zaslow, 2008). However, most children around the globe attend child care centers of mediocre quality (Peisner-Feinberg et al., 1999, Sylva, 2010; Tietze et al., 2013). In particular, vulnerable children, more precisely, children from ethnic minorities and children living in poverty, are more likely to attend early child care centers of lower quality, with less stimulating and engaging student-teacher interactions (Early et al., 2010; Espinosa, 2002; Gillanders et al., 2012). From a policy perspective, the large and consistent body of evidence clearly demonstrates the universal importance of high-quality child care for all children (Barnett & Frede, 2010). For the provision of high-quality services for children and their families, pro-

¹ The early childhood workforce is considered as the learner and recipient of professional development. Concerning the diversity of the workforce with respect to qualification and profession, the terms early childhood teacher, educator, professional, and practitioner are used interchangeably.

professional preparation and ongoing professional development for early childhood educators are essential (National Association for the Education of Young Children [NAEYC] & National Association of Child Care Resource & Referral Agencies [NACCRRA], 2011). Research indicates that child care quality is associated with teacher qualifications (Burchinal, Cryer, Clifford & Howes, 2002; Howes, 1997; NICHD, 2000b; 2002; Sonkoff & Philips, 2000). Over the decades, teacher qualifications have been measured with several indicators such as teachers' educational attainment and additional training (Burchinal, Hyson, & Zaslow, 2008; Kelley & Camilli, 2007). Early research indicated that having a Bachelor's degree clearly predicts teachers who provide high-quality child care experiences for children (Barnett, 2004; Burchinal et al., 2000; Helburn, 1995; Whitebook, 2002), while recent research queries those findings on the predictive potential of Bachelor's degrees (Early et al., 2007; Whitebook & Ryan, 2011). The meta-analysis by Kelley and Camilli (2007) showed a small correlational effect of Bachelor's degrees, but due to the correlational nature of the findings, it is possible "that any number of factors, besides having a Bachelor's degree, cause this effect" (p. 2). Besides methodological issues in early research (e.g., inconsistent definitions, inconsistent analytic methods, accounting for missing data and nested structures), Burchinal and colleagues (2008) explained the inconsistencies as due to the fact that teacher preparation programs were new at that time and the quality of the programs was uneven. Currently, findings suggest that the quality of child care is not necessarily predicted by teacher preparation and in particular by a Bachelor's degree (Institute of Medicine & National Research Council of the National Academies, 2012). High-quality child care experiences that enhance child development, defined as high-quality teacher-child interactions and the effective implementation of developmentally stimulating instructions and curricula, are not reliably produced by teacher preparation (Pianta et al., 2009). Several professional development programs exist that improve those aspects of child care quality (Fukkink & Lont, 2007; Pianta et al., 2009). Recent research reinforces the importance of ongoing professional development to support early childhood teachers in implementing specific practices (Burchinal et al., 2008). Several studies demonstrate the ability of in-service programs to improve observed quality in the areas of age-appropriate activities (Cassidy et al., 1995; Howe et al., 2011), instructional support, classroom management and caregiver responsiveness (Breffni, 2011; Wasik & Hindman, 2011) as well as language and literacy-specific classroom practices (Dickinson & Caswell, 2007; Neuman & Cunningham, 2009).

Beneficial outcomes for young children are anticipated when teachers receive specialized training (Fukkink & Lont, 2007). Through systematic in-service training for teachers, social abilities such as peer play, learning behavior, cooperation and positive behavior can be fostered (Gallagher, Abbott-Shim, & VandeWiele, 2011; Rhodes & Hennessy, 2000) and

challenging behavior can be reduced, i.e., negative affect, disruptive behavior, aggression (Stoiber & Gettinger, 2011; Snyder et al., 2011).

In contrast to the ample research base on the impact of in-service programs on social abilities, limited efforts have been made to examine the impact of professional development on early mathematics abilities. Findings suggest a positive impact of in-service training that supports teachers in facilitating individual learning opportunities and incorporating mathematical activities in daily routines on math abilities (Arnold, Fisher, Doctoroff, & Dobbs, 2002; Gallagher et al., 2011; Kopacsí & Hochwald, 1998). Similar results appear for some in-service programs fostering the language and literacy abilities of young children in center-based care. Training which focuses on interactions between teachers and children increase language productivity significantly (Girolametto, Weitzman, & Greenberg, 2003, Girolametto, Weitzman, Lefebvre, & Greenberg, 2007). Vocabulary growth is also achieved by in-service programs which encourage teachers to use questioning techniques and interaction strategies during book reading (Buschmann, Simon, Jooss, & Sachse, 2010b; Wasik, Bond, & Hindman, 2006). With a focus on aspects of literacy, for example, the study by O'Connor (1999) showed that the print awareness skills (e.g., letter naming or letter-word identification) could be fostered through professional development that helps teachers to implement "Ladders to Literacy" activities in classrooms. Children whose teachers received approximately 30 hours of training in interactive teaching, reading aloud and embedded literacy and writing activities during play as well as a set of books outperformed students whose teachers did not get training or material on early writing and reading abilities, including writing vocabulary, reading words, and hearing sounds in words (McGill-Franzen, Allington, Yokoi, & Brooks, 1999). Nevertheless, not all in-service programs have the potential to improve language and literacy development sustainably (Assel, Landry, Swank, & Gunnewig, 2007; Cusumano, Armstrong, Cohen, & Todd, 2006; Krause, Kofler, & Hofbauer, 2011). However, comparisons of different in-service treatments clearly indicated great variations in in-service training effectiveness due to intensity and amount (Shidler, 2009), delivery format (Lonigan, Farver, Philips, & Clancy-Menchetti, 2011), and target group of children (Krause, Kofler, & Hofbauer, 2011, O'Connor, 1999). The meta-analysis by Fukkink and Lont (2007) showed great variability in the effectiveness of training programs on caregiver competencies. It also implies that the potential benefits of professional development are predicted by training design aspects.

In summary, there is a consensus on the value of professional development to improve early child care practice and child development sustainably, but empirical findings on the extent to which in-service programs enhance early learning experiences in center-based care and the early development of young children are inconclusive. Questions remain

concerning which professional development input is most effective and for whom, for which outcomes, under what conditions. A systematization and aggregation of findings with correct calculations of effect magnitudes is needed to draw reliable and confident conclusions on the impacts and benefits of in-service professional development programs. Defining what is meant by in-service professional development is fundamental for the systematization of findings on in-service training effects. So far, an inconsistent use of the terms “professional development” or “training” occurs in the literature (Maxwell, Feild, & Clifford, 2006, Snyder, Hemmeter, Meeker, Kinder, Pasia, & Mclaughlin, 2012). In this investigation, in-service professional development is defined as learning and training opportunities, both off-site and onsite, for early childhood educators that work in center-based child care. In-service professional development can lead to some kind of qualification or certificate, but does not contribute to a formal college or university degree (e.g., Associate, Bachelor’s, or Master’s degree). Although, pre-service education and teacher preparation at college and university level is also known to predict early childhood educational practice and child development to a certain extent, this investigation concentrates exclusively on in-service programs.

1.1. Research objectives

There is a growing body of evidence that investigates whether in-service professional development makes a difference in quality ratings and child development or not. Since the meta-analysis by Fukkink and Lont (2007) that was limited to studies published before 2006, the volume of research literature has grown (Zaslow et al., 2010a; 2010b). Recent reviews show findings on in-service training or coaching effects on child development or quality ratings (cf. Isner et al., 2011; Tout, Isner, & Zaslow, 2011; Zaslow et al., 2010a; 2010b). However, these reviews are limited to the U.S. and restricted to journal articles, book chapters and governmental reports. Restrictions in the synthesis of findings in published scientific literature is a perpetual concern of publication bias. Published literature may be systematically different from unpublished documents due to selectivity from both researchers and journal publishers (cf. Sutton, 2004). In particular, it is widely acknowledged that “the dissemination of research findings is influenced by the nature and the direction of results” (Sterne, Egger, & Moher, 2008, p. 298). Further, restricting research syntheses to a territory or specific language may cause language bias, because authors are more likely to publish significant results from RCTs in English-language journals (Egger et al., 1997b). A limited number of studies indicates that calculations of intervention effects in meta-analyses differ when non-English-language trails are excluded (cf. Jüni,

Holenstein, Sterne, Bartlett, & Egger, 2002; Moher et al., 2003). Where there is controversial methodological discussion, most authors suggest the inclusion of studies published in other languages than English to increase precision and to prevent potential language bias (Jüni et al. 2002; Moher et al., 1996; 2000; 2003; Stern et al., 2008). Controversial, too, are the results from single studies on in-service training effects. The overview by Fukkink and Lont (2007) shows both positive and adverse effects of in-service training at caregiver level. Where findings from individual studies are inconsistent, it is possible to draw conclusions with more confidence with meta-analytic techniques (Camilli, Vargas & Yurecko, 2003, Kelley & Camilli, 2007). Meta-analysis provides a precise estimate of in-service training effects and explains heterogeneity between individual studies (Egger et al., 1997a). So far, meta-analyses are missing that quantitatively aggregate the growing body of research findings on in-service training effects in the early childhood field. Further, little is known about factors that influence in-service training effectiveness and the link between in-service training and student achievement.

Research objective 1: Precise estimate of in-service training effectiveness

The primary research objective of this project was to provide precise calculations of the impact of in-service training programs for preschool teachers on quality ratings and developmental outcomes for young children using several meta-analytic research procedures. First, systematic review procedures were carried out to bring together the findings on in-service training effects (see sections 2.2; 2.3; 2.4). A systematic review brings together the literature systematically on a specific question, using an explicit method to search the literature and conduct critical appraisals on the quality of studies. It allows the researcher to take the whole range of relevant findings from research on a particular topic into account (Akobeng, 2005a). Second, following the systematic review procedure, data from individual studies that provided sufficient statistical data were pooled and reanalyzed (see sections 4.1.2; 4.2.2; 4.3.2). The study results were converted into a standardized measure of effect size, measured as the mean difference in standard deviation units. Meta-analysis provides a precise effect size calculation and explains the heterogeneity in the results from the individual studies.

Research objective 2: Identification of in-service training approaches that work

Another research objective was to identify the most promising in-service training approaches to improve child care quality and/or to enhance student achievement. Drawing on evidence-based medicine (Akobeng, 2005b), evidence-based practice in education targets the improvement of teaching and interventions through the identification and promotion of practice that works and rejects those that are ineffective or adverse. Effect sizes

above 0.80 are considered as large, according to Cohen (1988). Using a meta-analytic procedure, the impacts of a single in-service program on several outcome measures of the same topic were aggregated to a treatment effect size. In-service training approaches with an aggregated treatment effect size of $g > 0.80$ were listed separately in a narrative review. The narrative review described and investigated key features of those programs in more detail than the meta-analysis.

Research objective 3: Identification of theoretical and methodological modifiers that influence in-service training effectiveness

The third research objective was the identification and investigation of theoretical and methodological factors that influence calculations of in-service training effects. Little is known about factors that influence in-service training effects in the early childhood field. In particular, there is a lack of information on “what is taught in teacher preparation and in-service programs, how that instruction is delivered, and the mechanism by which it translates into classroom practices” (Early et al., 2007, pp. 574-575). Findings from meta-analysis research demonstrate that there is a certain difficulty in the detection of treatment outcomes, because effect estimates are influenced by methodological and intervention features (Wilson & Lipsey, 2001). First, theoretical and methodological factors were identified through an extensive literature search and review. The review included findings from empirical studies and an intensive examination of suggestions from adult learning theory. Based on the review, a coding schema and theoretical framework were developed. The coding schema quantified information from relevant intervention studies on effect modifiers. The integrated framework illustrates several potential effect modifiers and the path from in-service professional training to child achievement. Second, potential effect modifiers that influence in-service training effect calculations were investigated through a moderator analysis. Moderator analysis is a statistical method that describes which factors significantly influence in-service training effects. Third, a qualitative approach was used that compares key features of highly effective programs. Effective programs are in-service treatments with a large aggregated treatment effect size ($g > 0.80$). The qualitative narrative review procedure offers additional insights on modifiers that substantially influence in-service training effects.

Research objective 4: Exploratory analysis of the path from in-service professional development to child outcomes moderated by in-service training effects on quality ratings.

The fourth research objective deals with the link from in-service programs to child outcomes. There are still very few publications available that delineate this path. In general, a two-step linkage is assumed. In-service training affects child care practice and quality.

Enhanced child care practice and quality lead to improvements in student achievement (cf. Yoon et al., 2007). This assumed linkage is incorporated in the integrated framework (see section 2.1.5). A fourth meta-analysis investigated the assumed path model. The meta-analysis consisted of studies that provide quantitative data on both quality ratings and child development (see section 4.4.2).

In summary, this research project describes the current state of research as well as the rigor of primary studies. Based on the findings from the meta-analysis, recommendations for practice and guidance for further research can be made. To a certain extent, theoretical models and frameworks of professional development can be evaluated and validated through moderator analysis.

1.2. Research procedure

The set of systematic reviews and meta-analyses contributes to the identification of effective approaches and methods that improve child care quality and the everyday child care experience of young children which lead to enhanced developmental outcomes. In order to gain an overview of relevant studies, research findings were analyzed with a systematic coding procedure to categorize the empirical findings from intervention studies. This systematization allowed for a reliable prediction of expected outcomes and the magnitude of in-service training impacts. Further, authoritative recommendations for whom and under what conditions in-service programs are effective can be made.

A five-step procedure was used to aggregate the research findings (see figure 1). The first step was a systematic and extended literature search to find studies related to the topic. The published and grey literature were searched in electronic data bases, reference lists and renowned journals.

The second step comprised a systematic coding procedure. In coding, hits from literature searches were screened regarding study quality and relevance. Studies that met quality standards completely or with reservations were examined using an extended coding schema. The extended coding schema included information on study methodology, in-service training, and contextual factors.

The third step consisted of a systematic review. Intervention studies with different evaluation designs (e.g., RCT, CRT, quasi-experimental, non-experimental or multiple-baseline/single-subject studies) published after 1970 were included. The aim of this procedure was to obtain an overview of studies measuring in-service training effects on quality ratings and child outcomes. The research objective was to investigate whether profes-

sional development is effective or not. In general, systematic reviews are dependent on primary research. Three systematic reviews of in-service training effects on 3 different outcomes were made: (1) quality ratings, (2) language and literacy development, and (3) social and behavioral skills. Following the systematic reviews, conclusions could be drawn using reported descriptive or inferential statistics. Based on the outcome-specific literature review, new research questions arose that could be answered through meta-analytic techniques.

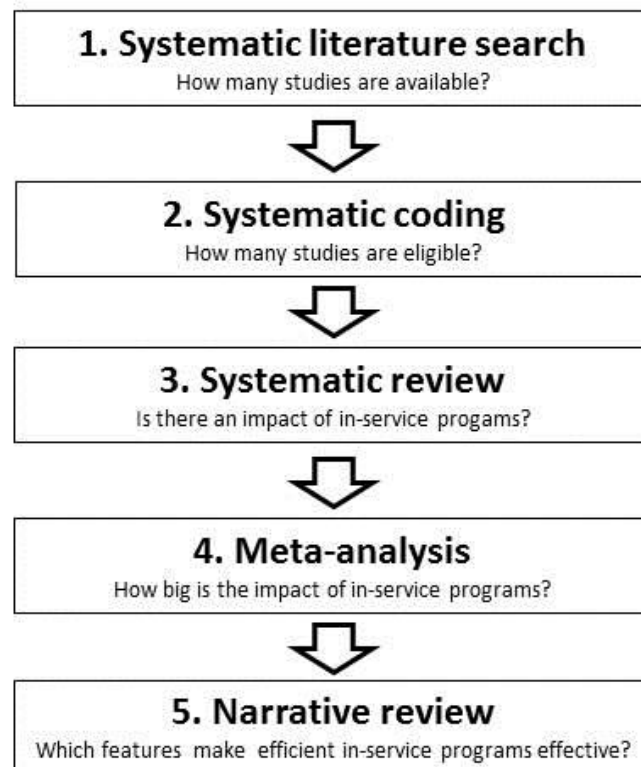


Figure 1. Five-step research procedure

The fourth step was a meta-analysis. For a meta-analysis, precise statistical data to calculate effect sizes are needed. Therefore, the meta-analyses included fewer studies than the systematic review. In general, meta-analyses increase statistical power (when data from studies are pooled) and the precision of treatment effect calculations (cf. Mulrow, 1994). Based on the available and eligible literature, meta-analyses were carried out for in-service training effects on the following outcomes: (1) quality ratings, (2) language and literacy development, and (3) social and behavioral skills. Another meta-analysis was realized on studies that provided data on both outcome categories, quality ratings and child development. Aggregated effect estimates were available for a summary effect size and aggregated effects per treatment. According to suggestions from Cohen (1988), in-service treatments with an aggregated effect size above 0.80 are considered as highly effective.

In the fifth step, highly effective in-service treatments were narratively investigated. In particular, key features of the effective in-service programs were examined through a qualitative content analysis. The analysis and categories were based on the available literature.

The dissertation is structured as followed: Section 2.1.1. delineates the path from in-service programs to child outcomes, mediated by changes in teacher performance and classroom quality. Sections 2.1.2.to 2.1.4. discuss theoretical assumptions and empirical findings on features that predict in-service training effectiveness. In section 2.1.5., a framework is presented that integrates the theoretical and empirical considerations to detect in-service training effects. In chapter 2.2., findings on the impact of in-service training on quality ratings are summarized in a systematic review. Another systematic review of in-service training effectiveness on language and literacy development is presented in chapter 2.3. The impact of professional development on the promotion of social skills and the reduction of challenging behavior is reported in chapter 2.4. The method section (chapter 3) provides a detailed overview of the methodology that was used to search the literature, code the studies, calculate effect sizes, and apply the meta-analyses and moderator analysis. Four meta-analyses are presented in the results section. Three meta-analyses investigated the potential of in-service programs to improve classroom quality (section 4.1.2.) and enhance child development (sections 4.2.2. and 4.3.2.). Section 4.4. consists of an exploratory examination of the linkage between in-service training and student achievement via quality improvements. Findings from the 4 meta-analyses and their potential effect modifiers and limitations are discussed in section 4.5. In the discussion section (chapter 5), the results are summarized and recommendations for professional development practice in the early childhood field as well as for further research in in-service training are described.

2. Theoretical framework and review of relevant literature

This dissertation is designed to examine the linkage between professional development, quality of early child care and student achievement. Chapter 2 consists of 4 sections. To answer the questions how professional development affects teacher, classroom, and student outcomes, several theoretical frameworks on professional development are examined. In particular, theoretical frameworks on the professional development process, training transfer factors and effective components of in-service training are delineated in the first section (see 2.1.). Findings from the literature are aggregated in an integrative framework to estimate the effects of professional development. Further, the question is raised whether it is generally possible to improve quality in early childhood education and to enhance student outcomes through in-service training. This question is answered through the systematization of research findings. A systematic review of the impact of quality ratings is shown in section 2.2. Research is reviewed on in-service training effects on language and literacy development in section 2.3., and section 2.4. aggregates the findings on the impact of in-service programs on social abilities and reduced problem behavior.

2.1. Theoretical frameworks

2.1.1. Process frameworks for professional development effects

Professional development, including pre-service and in-service, is viewed as the most promising means to ensure that all early childhood practitioners have the essential knowledge and skills to provide effective teaching and intervention practice (National Professional Development Center on Inclusion [NPDCI], 2008). Effective teaching and intervention practice is associated with the goal of improving child outcomes. Professional development can be defined as systematic efforts to ensure that early childhood practitioners are adequately qualified and effective in working with young children and their families (cf., Buysse, Winton, & Rous, 2009, NPDCI, 2008). Currently, professional development approaches, in particular the in-service efforts that are the major focus of this dissertation, range from traditional workshops and coursework with teachers absent from their classrooms to modern integrated approaches that provide intensive training with onsite support. Training sessions are offered by different providers with a range of qualifications and experiences. Further, professional development varies widely with respect to format (tradi-

tional vs. modern², individual vs. standardized, direct vs. indirect feedback) and content (Snyder et al., 2012). A special feature of the early childhood field is the heterogeneity of the early childhood practitioners who are widely diverse with respect to their professions, qualifications, experience and ethnicity (Buysse, Winton, & Rous, 2009). In-service professional development is defined as systematic efforts to facilitate learning for early childhood practitioners working in the field. It is designed to enhance practitioners' knowledge, skills, attitudes, awareness and performance that lead to high-quality learning experiences for young children in center-based care.

Professional development: a multiple step change process

One of the most robust and consistent findings in early childhood education research and literature is the association between high-quality child care and positive child outcomes (Burchinal, Kainz, & Cai, 2011; Mashburn et al., 2008). The STRUCTURE → PROCESS → OUTCOME framework of the NICHD (2002), describes direct and indirect effects of child care quality on the development of young children in center-based care. Quality of child care is measured in both process and structural features. While process features (e.g., teaching and interaction quality) are conjectured to have a direct impact on child development, structural variables (e.g., child-staff ratios, caregiver training) are assumed to have an indirect impact via process features (NICHD, 2002). It has become evident that a) process quality is linked to cognitive and social competences and b) the indirect path from structural features (mediated by process quality variables) to cognitive and social skills is significant.

Table 1. Frameworks on the process from professional development to child outcomes

Author/study	Target outcome teacher				Target outcome classroom	Target outcome children
	Orientation	Knowledge	Skills	Performance	Classroom quality	Child outcomes
Fukkink & Lont (2007)	✓	✓	✓	✓		✓
Hamre et al. (2012)	✓	✓	✓		✓	✓
Harvard Family Research Project (2006)		✓	✓	✓	✓	✓
Klein & Gomby (2008)				✓	✓	✓
Yoon et al. (2007)		✓	✓		✓	✓

Based on this paradigm, there are varieties of models that explain the path from caregiver training, including in-service and pre-service training, to improved teacher behavior, child care quality and/or enhanced child outcomes (Fukkink & Lont, 2007; Hamre et al., 2012; Harvard Family Research Project, 2006; Klein & Gomby, 2008; Yoon et al., 2007). Profes-

² Modern or reform-oriented traditional development incorporates more opportunities for in-depth, engaged learning activities (see Garet et al., 2001).

sional development affects student achievement through multiple steps and changes in a variety of outcomes (see table 1). It is assumed that changes appear at teacher level and/or at classroom level. These changes affect students' learning.

One of the most holistic frameworks to describe the process from professional development to beneficial child outcomes is the model of Fukkink and Lont (2007). In this framework, it is assumed that education and training of caregivers influence the professional competences of caregivers which underlie their interactional performance with children. The interactional performance of caregivers influences the development and behavior of children in center-based care (ibid., p. 297). The model distinguishes between caregiver knowledge, attitudes, and skills, whereby skills are assumed to be the most directly associated to process quality as the most visible part of caregiver competences (see figure 2). Fukkink and Lont (2007) found a fixed in-service training effect of $g = 0.45$ on outcomes at teacher level and a fixed effect of $g = 0.55$ for child behavior and development.

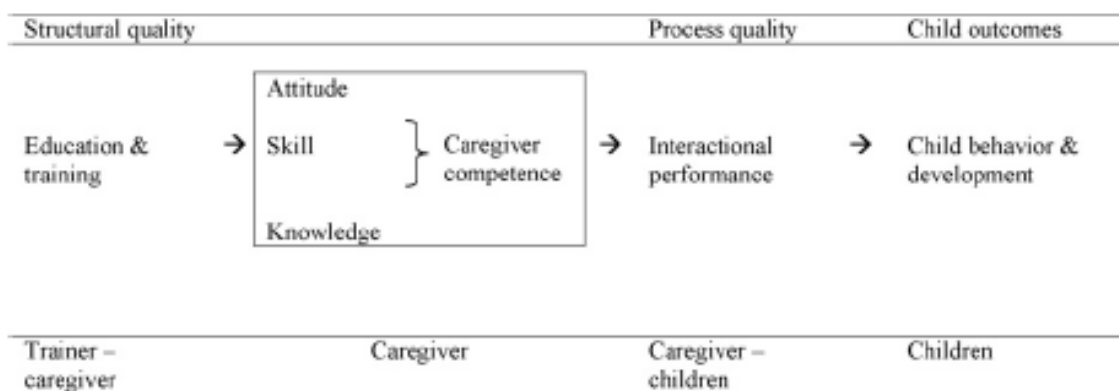


Figure 2. Model for teacher training effects (Fukkink & Lont, 2007, p. 297)

A very similar change model was used by Hamre et al. (2012) to delineate the process from coursework to effective teacher-child interactions during literacy and language activities that result in enhanced language and literacy abilities in young children. In their model, 4 steps were used to describe the process: professional development → teacher beliefs, knowledge and skills → classroom practice → child outcomes (ibid., p. 96). The findings show that a 14 weeks' course had a great impact in changing teacher beliefs ($ES = 0.43$ to 0.65) and knowledge ($ES = 0.49$ to 0.72) on language and literacy-related topics in comparison to a control group, but only modest evidence for the effect on some aspects of interactional quality. The model used in the Harvard Family Research Project (2006) links staff professional development to student achievement via multiple outcome steps. Better teacher knowledge, skills and competencies are viewed as short-term outcomes; improved practice as an intermediate outcome. Higher quality experiences are

considered as a long-term outcome that results in improved skills for children and youth (ibid., p. 3).

While other models integrate different target outcomes for teachers, the model of Klein and Gomby (2008) also emphasizes contextual and training factors that influence the process from professional development to child outcomes. Their framework was developed based on federal interventions and research that share the common assumption that professional development affects teacher practice in classrooms and that this practice carries over into beneficial child outcomes. Thus, early childhood teachers receive training to improve their practice and the overall quality in classrooms or/and to implement a new curriculum. Typically, in-service programs hire trainers/coaches to work with teachers initially or ongoing in an individual onsite format. According to Klein and Gomby's model, changes in teaching and instructional practice as well as quality improvements are direct results of training. Those changes continue into beneficial child outcomes. Klein and Gomby (2008) also presume that training for coaches/trainers improves the coaching/training delivered to teachers. Further, the effectiveness of training is moderated by workplace or teacher characteristics.

In the review on the effectiveness of teacher professional development by Yoon and colleagues (2007), a model was developed that describes the path how teacher training affects student achievement. According to their framework, professional development leads to enhanced teacher knowledge and skills that result in improved classroom teaching which affects student achievement. Additionally, contextual factors such as curricula, standards, accountability and assessment mediate the path from training to child outcomes and affect professional development (ibid., p 4).

In summary, most of the models describe a multiple step process to enhance student learning through in-service training of teachers. First, in-service training enhances teacher knowledge and skills and changes attitudes. Second, better knowledge and skills improve teaching and interventions as well as teacher-child interactions. Third, improved classroom quality enhances child development. However, there is only a limited amount of literature on theoretical models (see Fröhlich-Gildhoff, Nentwig-Gesemann, & Pietsch, 2011; Fried & Briedigkeit, 2008) and research findings that describe how professional development influences teacher knowledge and attitudes in the early childhood field (see Cunningham, 2007; Hamre et al., 2012; Pianta et al., 2014). Consequently, this dissertation concentrates on the linkage from in-service training to classroom quality and student achievement.

Some of the models consider contextual variables as factors that might influence the process. The workplace environment and organizational support as well as characteristics of trainer and trainee are considered as effect modifiers (Harvard Family Research Project,

2006; Klein & Gomby, 2008). These contextual variables are treated as important factors in several theoretical adult education models that influence the transfer of training to the workplace (Baldwin & Ford, 1988; Dubs, 1990; Rank & Wakenhut, 1998). Specific factors related to the early education field are curricula (Klein & Gomby, 2008; Yoon et al., 2007) and distal factors such as policy and standards (Harvard Family Research Project, 2006; Yoon et al., 2007).

2.1.2. Training transfer factors

The positive transfer of learning and training experiences to meaningful changes in work performance is a paramount concern in professional development efforts (Goldstein & Ford, 2002). "Positive transfer is defined as the degree to which trainees effectively apply the knowledge, skills, and attitudes gained in a training context to the job." (Baldwin & Ford, 1988, p. 63) Further, positive transfer appears when learned behavior is generalized to job situations and maintained throughout a period of time (ibid.). A transfer of knowledge, skills and attitudes is needed when the learning situation is temporally, contextually and spatially different from the applied work setting, situation or persons (Gnefkow, 2008; Staehle, Conrad, & Sydow, 1999). Transfer distance, in terms of near and far transfer, is discussed as a decisive factor that predicts positive transfer (Barnett & Ceci, 2002). In particular, it is assumed that highly similar learning tasks (near transfer) are more likely to be generalized and transferred into workplace performance (Blume et al., 2010; Joyce & Showers, 1981). Barnett and Ceci (2002) developed a taxonomy of near and far transfer conditions. They distinguish between a) content (that is transferred in terms of its specificity, generalizability, nature of performance change and memory demands) and b) transfer context (when and where the transfer takes place, regarding physical, temporal, functional and social contextual issues as well as knowledge domains).

There are several theoretical models of training transfer factors (Baldwin & Ford, 1988; Dubs, 1990; Huczynski & Lewis, 1980; Rank & Wakenhut, 1998) and a number of authors have made efforts to summarize the empirical findings on positive transfer predictors (Baldwin & Ford, 1988; Baldwin, Ford, & Blume, 2009; Blume et al., 2010). In the literature, training design, trainee characteristics, and the work environment are mechanisms responsible for the transfer of training (Baldwin & Ford, 1988; Dubs, 1990; Huczynski & Lewis, 1980; Rank & Wakenhut, 1998). *Professional development design* is responsible for the extent to which trainees perform differently once back on the job. Alongside process and component factors, the content as well as sequencing, the amount, intensity, and duration of professional development and principles of learning are related to the ex-

tent of training transfer. *Trainee characteristics*, in particular abilities, motivation, and personality aspects, are related to successful transfer. Additionally, voluntary participation and strong beliefs that training leads to improvements in job performance are discussed as moderating factors. Further, the *work environment*, including support by heads of institutions, opportunities to use new skills and issues of company climate are regarded as important predictors. The transfer trials that participants attempt in workplace situations are thematized as an additional factor by Rank and Wakenhut (1998). In teacher education, the amount of practice on the original task is assumed to be associated with positive transfer (Joyce & Showers, 1981).

Over the last decades, several empirical studies have investigated learning and training transfer factors on a variety of outcomes such as knowledge, motivation, post-training self-efficacy, training performance or transfer performance (cf. Alvarez, Salas, Garofoano, 2004; Blume et al., 2010; Burke & Hutchins, 2007; Cheng & Ho, 2001; Colquitt, LePine, & Noe, 2000). However, the evidence base for moderators which could influence the transfer from training to the application of new skills (in the literature defined as use, workplace performance, training effectiveness or transfer performance) in the workplace is far smaller. So far, only a few significant moderators have appeared to be significantly related to post-training performance at the workplace; these are listed in table 2.

Table 2. Evidence base on training transfer factors to change workplace performance

Training design	Trainee characteristics	Work environment
• Learning principles • Learning goals • Content relevance • Practice & feedback • Behavior modeling • Post-training intervention • High difficulty • Active learning	• Self-efficacy • Motivation • Locus of control • Knowledge • Anxiety • Cognitive ability • Skill acquisition • Reaction to training	• Positive transfer climate/environment • Role of supervisor/peers • Organizational commitment • Career planning

Training design

In their systematic review, including studies from 1993 to 2002, Alvarez and colleagues (2004) conclude that “*learning principles* such as practice, part- versus whole-task learning, and feedback were positively associated with transfer performance” (p. 405). Further, well-defined *learning goals* and training objectives are considered to increase training transfer (Locke, Shaw, Saari, & Latham, 1981; Richman-Hirsch, 2001; Taylor, Russ-Eft, & Chan, 2005; Wexley & Baldwin, 1986; Wexley & Nemeroff, 1975). In particular, both participative and assigned goals, in terms of proximal (short-term) and distal goals, are associated with higher values of training transfer (cf. Brown, 2005; Taylor et al., 2005). When learning goals and objectives are transparent, trainees are more likely to transfer new skills when they have a clear understanding of the behavior that is required after training

(Kontoghiorghes, 2001). Burke and Hutchins (2007) suggest that "trainees must see a close relationship between training content and work tasks to transfer skills to the work setting" (p. 274). Strong correlations between *content relevance* and transfer outcomes were found (Lim & Morris, 2006; Rodriguez & Gregory, 2005). Moreover, *behavioral practice* and mental rehearsal during professional development is related to positive transfer (Ford & Kraiger, 1995; Warr & Allan, 1998). Lee and Kahnweiler (2000) found significantly higher transfer to work tasks when trainees received *feedback*, reinforcement and remediation opportunities. For the generalization of novel tasks, *behavioral modeling* strategies, including rule codes as well as descriptive and rule-oriented learning points, were statistically related to post-training behavior (Decker, 1980; 1982; Decker & Nathan, 1985). The meta-analysis by Taylor et al. (2005) clearly indicates that behavioral modeling strategies have a greater impact on transfer when positive and negative models are used. Some empirical findings from training for pilots and firefighters indicate that error-based examples or detailed case studies on what can go wrong enhance transfer performance (Ivancic & Hesketh, 2000; Smith-Jentsch, Salas, & Brannick, 2001). For the maintenance of new skills, systematic *post-training interventions (follow-up)* influenced transfer performance positive (Burke, 1997; Gist & Stevens, 1998; Krijger & Pol, 1995; Morin & Latham, 2000; Richman-Hirsch, 2001). Post-training interventions consist of "goal setting, visualization exercises, and mastery orientation manipulations" (Alvarez et al., 2004, p. 405). Additionally, *high difficulty levels* of training, that require trainees to learn tasks that are more difficult than those at the workplace, were also found to be significantly related to transfer performance (Doane et al., 1996; 1999; Gist & Stevens, 1998). *Active learning methods* (i.e., behavior modeling, dialogue, feedback) are assumed to be more effective in transferring performance compared to passive, traditional formats such as lectures and workshops (Burke & Hutchins, 2007). The meta-analysis by Burke et al. (2006) on worker safety and health training indicates that more engaging training methods with active participation by trainees increase knowledge acquisition and safety performance and reduce negative health and safety outcomes such as injuries.

Trainee characteristics

Several studies confirm that *self-efficacy* is positively related to post-training behavior, transfer performance and skill maintenance (cf. Alvarez et al., 2004; Bell & Kozlowski, 2002; Burke & Hutchins, 2007; Cheng & Ho, 2001; Chiaburu & Marinova, 2005; Colquitt, LePine, & Noe, 2000; Ford et al., 1998; Gaudine & Saks, 2004; Gist, 1989; Gist, Stevens, & Bavetta, 1991; Mathieu, Tannenbaum, & Salas, 1992; Saks, 1995; Stevens & Gist, 1997; Tannenbaum et al., 1991; Warr, Allan, & Birdi, 1999). Modest relationships between trainee *motivation* and transfer performance (determined as use or effectiveness) were

found when they were measured objectively or by others instead of by self-evaluation (Blume et al., 2010). Further, trainees with high levels of (internal) *locus of control* were more likely to transfer newly acquired skills to the workplace (Cheng & Ho, 2001; Colquitt, LePine, & Noe, 2000; Tziner, Haccoun, & Kadish, 1991).

In the meta-analysis by Blume and colleagues (2010), small correlations between *post-training knowledge* and the use of new skills ($r = .20$) and training effectiveness ($r = .17$) were found when measured externally or objectively. This finding is compatible and coherent with models of change in teacher behavior. A small body of evidence indicates that *cognitive ability* (Kanfer & Ackerman, 1989; Robertson & Downs, 1979), *anxiety*, *skill acquisition* during training and *reactions* towards training are related to trainee performance and training effectiveness (Colquitt, LePine, & Noe, 2000).

Work Environment

Several studies confirm the relationship between the transfer performance of newly acquired skills to the workplace and a positive *transfer environment/climate*, a climate that encourages trainee to implement new skills (Alvarez et al., 2004; Blume et al., 2010; Colquitt, LePine, & Noe, 2000; Fecteau et al., 1995; Rouiller & Goldstein, 1993; Rynes & Rosen, 1995; Tracey, Tannenbaum, & Kavanagh, 1995). *Support by supervisors and peers* is considered as an important moderator to facilitate the transfer of training and influence the use of trained skills at the workplace (Burke & Hutchins, 2007; Colquitt, LePin, & Noe, 2000). Further, trainees who are highly committed to the organization, individually involved and identify with the organization (*organizational commitment*) are more likely to demonstrate high transfer performance. Additionally, *career planning*, the extent to which trainees are committed to a specific plan for achieving career goals, is related to job performance after training.

In summary, in the research on training transfer, several factors have been scientifically identified that moderate the effectiveness of in-service programs and result in meaningful changes in work performance. This empirical knowledge base needs to be compared with the literature on professional development effectiveness in early childhood education.

2.1.3. Professional development components and potential effect modifiers

The framework by Buysse, Winton and Rous (2009) provides a broader overview of components that are assumed to make professional development effective in improving classroom quality and child outcomes. It helps to design, implement and evaluate professional development in the early childhood field (see figure 3). By definition, professional devel-

opment encompasses all types of facilitated learning opportunities to support the acquisition of professional knowledge, skills, and attitudes and results in improved teaching and intervention practice and beneficial child outcomes. According to the authors, professional development integrates 3 intersecting core components (the “who”, the “what” and the “how”) that are surrounded by contextual factors such as organizational structure, resources, policies, evaluation as well as access and outreach.

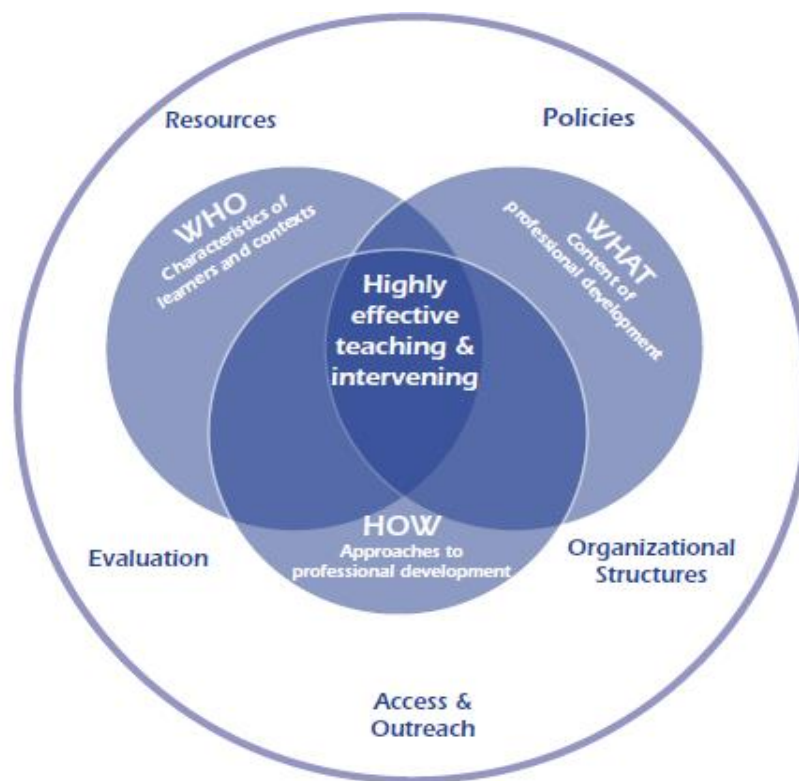


Figure 3. Conceptual framework for professional development (Buysse, Winton, & Rous, 2009, p. 239)

The “how” component encompasses the organization and facilitation of the learning experience and describes delivery formats for professional development, e.g., approaches, models, methods used to support self-directed, experientially oriented learning that is relevant to practice (Buysse, Winton, Rous, 2009; National Professional Development Center on Inclusion [NPDCI], 2008). The “how” includes features such as amount, duration, intensity, format, grouping, active learning, and the provision of feedback.

The “what” component determines specific knowledge, skills, and attitudes that are emphasized by professional development programs. Professional development should focus on professional practice (e.g., evidence-based instructional strategies, a high-quality classroom environment) and consists of content-specific instructions. These content-specific instructions should be coherent with the curriculum material, learning standards and instructional goals that early childhood educators use in practice (Birman et al., 2000;

Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Elmore, 2002; Garet et al., 2001; Yoon et al., 2007; Zaslow et al., 2010a).

The “who” component emphasizes learners and providers of professional development. In early childhood education, practitioners, as well as the children and families they serve, vary widely with respect to their profession, qualifications, experience, culture and ethnicity. Therefore, professional development has to consider the organizational context and the heterogeneity of learners and providers (Buysee, Winton, & Rous, 2009).

The framework provides a brief overview of features and components for the evaluation of in-service training effects on teacher, classroom, and student outcomes. Based on this framework, Snyder et al. (2012) carried out a narrative review with 256 studies reporting empirical evidence on professional development programs for early childhood educators. The review demonstrated that the framework provides a clear consensus of features to systematize professional development interventions. In consequence, the framework by Buysse, Winton and Rous (2009) has guided the investigations of the moderators in this dissertation.

Several other theoretical thoughts and literature reviews were conducted to identify the effective features of professional development (cf., Blank & de las Alas, 2009; Elmore, 2002; Garet et al., 2001; Ingvarson, Meiers, & Beavis, 2005; Yoon et al., 2007; Zaslow et al., 2010a). Initial conclusions from the literature and empirical findings from early education research for the design of effective professional development programs are described below. Other potential effect modifiers are trainee characteristics, contextual features, and program fidelity.

1. Duration, amount and intensity of professional development

It is widely recommended that the duration and amount of professional development should be matched to the content being conveyed. It should provide appropriate time for active learning that is matched to the experience of teachers and contextual factors of their centers/schools. Further, professional development should offer learning opportunities that are intensive and sustained over time to ensure continuous improvements (Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Elmore, 2002; Garet et al., 2001; Ingvarson et al., 2005; Isner et al., 2011; Yoon et al., 2007; Zaslow et al., 2010a).

The findings from (quasi-)experimental studies on a “golden” rule for training duration are inconsistent. On the one hand, several studies demonstrate that intensive and sustained training formats with a duration of 12 or more months result in improved classroom quality (Algozzine et al., 2011; Grace et al., 2008; Roehrig et al., 2011; Sheridan, Clarke, Knoche, & Edwards, 2006) and beneficial child outcomes (Kopacsi & Hochwald, 1998; Mashburn et al., 2010; Mohler et al., 2009; Shidler, 2009). On the other hand, short-term in-service

programs also result in equally beneficial effects on general or single aspects of classroom quality (Breffni, 2011; Cunningham, 2007; Englund, 2010; Howe et al., 2011; Neuman & Cunningham, 2009; Neuman & Wright, 2010) as well as on developmental outcomes for children (Girolametto et al., 2003; 2007; Graul & Zeece, 1990; Hsieh, 2005; McGill-Franzen et al., 1999; Onchwari & Keengwe, 2010). Examining the differential effects of interventions of one or 2 years of coaching, findings show that multiple years of intensive coaching are linked to stronger impacts on interactional quality in early childhood classrooms and more enhanced developmental (language and literacy) gains for young children (Hindman & Wasik, 2012; Landry et al.; 2006, 2009). The developmental gains for young children in teachers' second year of training can be considered as the results of successful implementation of improved child care practice and ongoing support to maintain the target practice. To date, a limited number of studies investigates the retention and maintenance of successfully implemented practice through follow-up measures (e.g., Englund, 2010; Simon & Sachse, 2013).

2. *Delivery format of professional development*

The literature emphasizes the importance of the form of delivery of professional development. Traditional approaches (e.g., workshops or courses) are considered as less effective in fostering meaningful changes in classroom practice than modern approaches that are thought to be more responsive to teacher learning (Birman et al., 2000; Garet et al., 2001). In summary, 5 in-service formats are used predominantly in in-service programs for early childhood educators and can be differentiated as (1) traditional (isolated) workshops and initial training institutes, (2) coursework with multiple, ongoing and consecutive sessions, (3) remote approaches with web- and satellite based resources, (4) communities of practice or learning with shared inquiry and enterprise, and (5) ongoing, onsite training either delivered in-class or distally.

A small but growing body of evaluation studies validates the contribution of *delivery format* and *in-service models* on training effectiveness in the early childhood field. Several studies, experimenting with delivery formats and approaches, provide substantial evidence that there are indeed *differential effects of delivery mode* on quality rating and child development (Assel et al., 2007; Downer et al., 2011b; Girolametto, et al., 2003; 2004; 2007; Howe et al., 2011; Japel, 2009; Lonigan et al., 2011; O'Connor, 1999).

The current evidence base supports the assumption that *onsite formats*, either as the sole intervention or in addition to other delivery modes, are more strongly related to beneficial classroom and child outcomes than traditional delivery types such as courses or workshops (Howe et al., 2011; Japel, 2009; Lonigan et al., 2011; Neuman & Wright, 2010). Several onsite formats are used in early education teacher training (e.g., technical assis-

tance, mentoring, or consultation). In particular, *coaching* is an established format to change teacher behavior (Hsieh, 2009), implement new curricula or instructional strategies with high fidelity (Armstrong, 2008), and to improve process and interaction quality (Cunningham, 2007; Hindman & Wasik, 2012; Neuman & Cunningham, 2009; Neuman & Wright, 2010; Powell et al., 2010). In addition to quality, coaching increased the frequency of the use of evidence-based strategies. The findings by Rudd et al. (2009), using multiple measures during the training phase, show that the use of math-mediated language was more frequent during the coaching phase than during the workshop phase. Coaching teachers is also known to produce beneficial developmental outcomes for young children (Hindman & Wasik, 2012; Hsieh, 2009; Mohler et al., 2009; Powell et al., 2010; Shidler, 2009). In the research synthesis on 9 coaching studies, Gupta and Daniels (2012) remark that there is evidence for the effectiveness of coaching on several outcomes, but less is known about the coaching model itself, because most of the studies do not describe the behavior and actions that the coaches used to support teachers. Powell et al. (2010) compared the impact of remote (distance video feedback) and onsite coaching conditions on language-specific classroom quality and the language and literacy skills of preschoolers. However, both coaching conditions resulted in considerable effects with no significant difference of the 2 coaching conditions (ibid.).

3. Practice-focused professional development

The literature suggests that professional development should be based on adult learning theories (Elmore, 2002; Yoon et al., 2007) and that in-service programs should have an explicit focus on the strengthening of early child care practice, rather than teacher knowledge (Buysse, Winton, & Rous, 2009; Zaslow et al., 2010a). According to Zaslow et al. (2010a), the evidence base on the effectiveness of practice-focused professional development is inconsistent, as it compares studies using modeling, feedback and onsite support with or without traditional coursework. Snyder and colleagues (2012) identified “assignments” and “back-home plans” as frequently used follow-up methods in which trainees are encouraged to accomplish professional development-related exercises in their classrooms after training.

A growing body of research evaluates the impact of in-service programs using *performance-based assignments* (Buschmann, Jooss, Simon, & Sachse, 2010a; Buschmann, Simon, Jooss, & Sachse, 2010b; Dickinson & Caswell, 2007; Girolametto et al., 2003; 2012). Performance-based assignments are concrete instructions for exercises for implementing new skills and knowledge at the workplace. Adult learning theory suggests that in-service programs are more effective when specific skill practice is included and when learning takes place in an authentic context (Girolametto et al., 2012; Landry, 2006). Per-

formance-based assignments require trainees to perform a specific piece of classroom practice, instructional strategy, or learning activity (Dickinson & Caswell, 2007). Trainees are also required to reflect on their own behavior, child responses, and conversations with children by documenting them using audio and videotaping as well as transcription. These documented efforts to implement novel skills and practices in early childhood classrooms can be analyzed by self-reflection, feedback by a coach/supervisor (Dickinson & Caswell, 2007; Girolametto et al., 2003; 2012) or group feedback (Buschmann et al., 2010a; 2010b).

4. *Feedback*

It is widely recognized that feedback and guidance are promising approaches to help teachers apply new skills in their classroom practice. Performance feedback guides teachers on how to apply specific practice. This is done through coaching, mentoring, consulting or technical assistance (Buysse, Winton, & Rous, 2009; Elmore, 2002; Ingvarson et al., 2005).

A few studies have investigated the impact of *data-driven feedback* formats. In-service programs use intervention fidelity checklists (cf. Brown, 2008; Jo Rodriguez et al., 2009), teacher behavior observations (Landry et al., 2009), and progress monitoring scores of student outcomes (Grace et al., 2008; Landry et al., 2009; 2011; Tyler, 2009) to provide data-driven feedback. Findings from a CIRCLE project that compared 4 interventions demonstrate that the use of progress monitoring, in particular via a technology-driven progress monitoring tool, resulted in robust improvements in the quality and quantity of teacher instructions and children's outcomes (Landry et al., 2009).

Child assessment can be used for the ongoing monitoring of professional development to provide distal feedback. Professional development that targets student learning should be evaluated consistently with student outcomes (Elmore, 2002; Ingvarson et al., 2005; Landry, 2005; Zaslow et al., 2010a). The study by Landry et al. (2009) demonstrates improvements in teaching practice and student achievement from a PDA-based progress monitoring tool that links student learning to new, applied strategies and skills. The *student progress monitoring tool* a) informs teachers immediately about individual learning needs and the developmental progress of children in their classrooms, b) offers information to choose supporting practices and strategies that are needed and c) provides feedback that allows for individual, in-service adjustment based on student achievement (Landry et al., 2011).

5. *Content-focused professional development*

The literature recommends that in-service approaches and activities should be content-focused and should target a specific subject area or teaching strategy rather than general instruction (Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Elmore, 2002; Garet et al., 2001; Ingvarson et al., 2005; Landry, 2005; Yoon et al., 2007; Zaslow et al., 2010a). Indeed, the effects of professional development on early educator practice were found to be larger when the content was more specific (Fukkink & Lont, 2007). Further, it is proposed that observational measures can help professional development to provide articulated and specific goals and guide in-service interventions in areas of practice that are relevant to beneficial child outcomes (Landry, 2005; Zaslow et al., 2010a).

Some findings support the assumption that the use of performance and environmental *checklists as benchmarks* is a valuable tool. Checklists promote early educators' understanding, adoption, and use of practices, and strengthen the benefits of coaching procedures (Cordell, 2010; Dunst & Raab, 2010; Trivette et al., 2012a). Moreover, checklists or observation forms contain evidence-based practices that operationalize training objectives. The scales provide targets and content for professional development activities. Several in-service programs apply *scale-based approaches* grounded on external quality rating scales (Japel, 2009; Manningham, 2008; McNerney, Nielson, & Clay, 2006), and teacher behavior observation forms (Beller, Merkens, & Preissing, 2007; Beller & Beller, 2009). Professional development based on observation tools that consider instructional practice and the classroom environment can a) guide decisions about material purchases in classrooms, b) adjust in-service sequencing and delivery and c) guide coaches in their one-on-one work with teachers (McNerney et al., 2006).

6. *Adult learning principles*

There is a consensus that professional development programs that provide *active learning* opportunities enhance teacher knowledge and skills and improve classroom practice. Teachers are actively engaged in their own learning through meaningful discussions, planning and exercises as part of the program (Birman et al., 2000; Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Garet et al., 2001; Ingvarson et al., 2005). Four adult learning methods characterize active learning: practicing, evaluation, reflection, and mastery. Programs that use those active learning strategies have a stronger impact on the learning outcomes of teachers and the use of adaptive strategies in special education than methods which merely introduce new information or illustrate and demonstrate new knowledge (Dunst & Trivette, 2011; Dunst & Raab, 2010; Dunst, Trivette, & Hamby, 2010; Trivette, Raab, & Dunst, 2012a; 2012b).

The literature suggests that professional development goals need to be *coherent with the policies, standards, curricula, and assessment* used in early childhood education. To ensure the application of new skills and techniques, professional development needs to be aligned within organizational structures, instructional goals, and the curriculum material that early education teachers use. In-service training should be consistent with teachers' learning and experience, build on teachers' goals and earlier activities, provide follow-up activities and involve teachers in reflection and discussion (Birman et al., 2000; Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Elmore, 2002; Garet et al., 2001; Yoon et al., 2007; Zaslow et al., 2010a).

It is assumed that the *collective participation* of teachers from the same classroom or center/school is more effective, because they are more likely to integrate new knowledge and skills. In particular, the sharing of curriculum material and assessment requirements and the common understanding of goals, methods, problems and solutions are more likely to ensure the sustainability of new techniques and skills. Further, collective and joint participation can contribute to a shared professional culture within the center/school (Birman et al., 2000; Blank & de las Alas, 2009; Elmore, 2002; Garet et al., 2001; Ingvarson et al., 2005; Zaslow et al., 2010a).

At this stage, there is no evidence available that confirms these assumptions for the early education field. In consequence, active learning, coherence, and collective participation are exploratory research objectives that were included in the coding procedure.

7. *Fidelity*

The results from Trivette et al. (2012b) demonstrate that *in-service intervention fidelity and quality of delivery* matter. In particular, effects were larger when coaches addressed all aspects of the Participatory Adult Learning Strategy (PALS) used in the coaching procedure. The effect was significantly moderated through PALS-fidelity during coaching. Significant standardized mean differences on CLASS-scores ($d = 0.32$ to 0.67) were found among trainees that received coaching with a higher degree of PALS strategies and for those with low fidelity. The level and amount of participation as fidelity index did not moderate the use of new (mathematic) strategies and quality of instruction (Varol et al., 2012b). Trivette et al. (2012b) found considerable variations in intervention fidelity within one coach. In contrast, in the large-scale study by LoCasale-Crouch et al. (2011), the observed quality of delivery, fidelity of the NCRECE course and confidence in instructors was stable. The authors kept the conditions stable through ongoing support for instructors and pre-assembled course material for each session (e.g., power point, video ratings). However, where in-service programs are inadequately implemented, it is almost impossible to detect professional development effects (Yoon et al., 2007).

These findings suggest that the quality of professional development models matter. In addition to in-service program fidelity and adherence, *teacher fidelity* matters. *Teacher or curriculum fidelity* is the extent to which teachers implement new skills and/or instructional practice in classrooms after participating in in-service training. Curriculum fidelity - the quality of program and curriculum implementation – is widely considered as a relevant outcome of in-service education and a predictor for effects on child outcomes. Curriculum fidelity is the extent to which a curriculum is implemented as intended (Dane & Schneider, 1998). In general, higher fidelity scores are suggested to be related to better child outcomes (Domitrovich et al., 2010; Hamre et al., 2010). Further, professional development effects measured by quality ratings (Justice et al., 2008) and child outcomes (Jo Rodriguez, Loman, & Horner, 2009; Jungmann, Koch, & Etzien, 2013) were moderated to the degree of fidelity with which teachers implemented intended strategies.

8. *Trainee characteristics and background conditions*

Teacher and classroom characteristics (e.g., years of education, ethnicity, teacher beliefs, anger and anxiety scores, and the classroom environment) predict the responsiveness of trainees to coursework or consultancy and their attendance levels (Hatfield et al., 2012). Further, there is great variation in the degree to which early childhood teachers implement new teaching strategies after in-service training, even within 5 teachers (Hsieh et al., 2009). Given these facts, it is easy to conclude that there is a need for the investigation of trainee characteristics and aspects of their backgrounds as potential moderators of in-service training effectiveness. In early education research, the *initial score level* has been discussed. Landry et al. (2009) demonstrate that professional development is particularly beneficial for teachers who initially perform poorly or provide at least average teaching quality. In contrast, Hindman and Wasik (2012) show that teachers at all initial performance levels made equivalent gains in interactional quality and the language and literacy environment.

Focusing on the *educational background of teachers*, the findings are inconsistent. The study by Landry and colleagues (2006) clearly indicates significant positive effects on student gains in phonological awareness, receptive language and alphabetic knowledge for the first training year for all teachers. However, the in-service training effect on phonological awareness emerged only in the second year of training when teachers had more years of pre-service education (≥ 4 years conform to BA or MA vs. ≤ 2 years conform to AA, CDA or less). Hamre et al. (2012) found that their 14-session course “was equally effective across teachers with less than an associate’s degree as well as those with advanced degrees” (p. 88) for quality scores on the CLASS.

Further, contextual variables such as *child care system/provider* seem to matter. Public school educators showed a higher average number of math activities and higher quality of math instruction than Head Start teachers did after training (Varol et al., 2012a; 2012b). Requests for workshops, mentoring and online resources differed between Head Start and public school educators, indicating variation in training needs due to different classroom experiences. Hamre and colleagues (2012) could not find significant posttest differences in quality scores on the CLASS between teachers working in Head Start centers or other early childhood programs. Nevertheless, *child care center effects* on in-service training gains were found for children's language and literacy scores with high variations in the areas of receptive language and auditory comprehension (Landry et al., 2006). Further, there is some evidence that *program length and the operating hours of the center* moderate the effects of in-service programs on the language and literacy scores of preschoolers when half vs. full day programs are compared (ibid.). Within centers, the *dominant spoken language in classrooms* seems to prejudice the effectiveness of programs. Some variations were found in language and literacy growth among students between classrooms in which English was the dominant language compared to classrooms in which Spanish or other languages were spoken predominantly (Downer, Lopez, Grimm, Hamagami, Pianta, & Howes, 2011a). Notably, in-service training effects were found for classrooms with English as the main spoken language and that received high levels of support (ibid., p. 189). In addition to center and classroom effects, in the multi-state CIRCLE professional development implementation study the in-service training impact on the receptive language scores of preschoolers was moderated by *federal state* (Landry et al., 2009). The effects on teacher behavior outcomes were similar across the 4 sites. The authors presume that variations appeared due to differences in teachers, children, and classroom characteristics, because Maryland and Florida vary essentially in the number of second language learners and required teacher qualification levels. In summary, the findings demonstrate that in-service programs can improve early childhood practice across diverse settings, but that specific characteristics of teachers and center issues might inhibit the full potential of in-service programs. These aspects need to be considered as potential moderators in further professional development interventions and research.

9. *Curricula as external and internal moderators*

There are several studies that evaluate approaches that use *curriculum implementation* to enhance interaction-specific (Pianta et al., 2008) or domain-specific quality (Algozzine et al., 2011; Clancy-Menchetti, 2006; Grace et al., 2008; Lonigan et al., 2011; Sibley & Sewell, 2011a; Sprague et al., 2009). A large number of in-service programs bank on the implementation of social-emotional curricula (cf. Baker-Henningham et al., 2012; Cordell,

2010; Domitrovich, Cortes & Greenberg, 2007; Lloyd & Millenky, 2011; Morris et al., 2010; Shernoff & Kratochwill, 2007), language and literacy-specific curricula or programs (cf. Assel, Landry, Swank, & Gunnewig, 2007; Lonigan et al., 2011; Mashburn et al., 2010) and/or a set of domain-specific learning activities (Buysse, Castro, & Peisner-Feinberg, 2010; Powell et al., 2010; Landry et al., 2006). Klein and Gomby (2008) mention that the implementation of a new curriculum might be confounded with the in-service training interventions. This makes it difficult to attribute the impact clearly to professional development. Findings by Landry et al. (2006) show that the *type of curriculum* used by child care sites, both domain-specific and universal, moderates the effectiveness of training programs aimed at language outcomes. Therefore, it is necessary to control for the presence of universal and domain-specific curricula to be able to reliably attribute effects to professional development programs.

To sum up, the review of the literature illuminates key features of in-service programs and contextual variables that are suggested to moderate in-service training effects on quality ratings and student achievement. However, most of the mentioned features are based on single studies. Further research and meta-analytic techniques are needed to investigate factors that explain the great variation in in-service training effectiveness. In addition to professional development features and training transfer factors, the methodological quality of studies is known to moderate the magnitude of effect sizes.

2.1.4. Methodological quality and measurement features moderating in-service training effects

This section primarily focuses on the rigor and methodological quality of studies that are needed to detect the impact of professional development on quality ratings and child outcomes. A poorly designed evaluation or inadequately accomplished training program makes it difficult to identify the effects of teacher training (Yoon et al., 2007). Strong empirical evidence can be assumed when studies apply research standards correctly (What Works Clearinghouse [WWC], 2011). The link between professional development and student achievement is confounded by aspects of study design, measurement features, and effect size calculations.

Study design

A rigorous study design has to ensure the internal validity of causal inference, external validity, and have adequate statistical power to detect the effects of in-service programs

(National Research Council, 2004a; Yoon et al., 2007). “Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations ” (WWC, 2011, p. 11) In particular, internal validity is needed to eliminate competing factors and explanations (e.g., curricula, materials) for the effects of in-service programs on classroom and student improvements (Yoon et al., 2007). Studies with (quasi-)experimental designs are warranted to measure the impact of professional development on students and quality gains. Further, study designs must be accomplished with high fidelity (ibid.) and levels of evidence can be categorized by randomization, attrition and equivalence (WWC, 2011). Randomization can be taken to mean that study participants (students, teachers, classrooms, schools) are assigned randomly (by chance) either to intervention or control conditions. The unit of analysis “should match the number of ‘units’ that were randomized” (Deeks, Higgins, & Altman, 2008, p. 260). More precisely, units of analysis can be biased a) when groups of individuals (e.g., classrooms, centers) are randomized in the same condition (i.e., cluster-randomized trials), b) participants receive a sequence of interventions (i.e., cross-over trials, multiple treatments) or c) repeated measures are made (ibid.). In quasi-experimental designs, study participants are not randomly assigned and therefore QEDs must demonstrate the equivalence of intervention and control groups. High-quality RCTs also provide information on the equivalence of groups. Further, attrition rate is an important indicator for levels of evidence and the quality of studies. High attrition rates compromise the initial equivalence of groups and bias effect sizes (cf. WWC, 2011).

Measurement features

Measures must be valid, reliable, and age-appropriate (WWC, 2011, Yoon et al., 2007). The *psychometric properties* of outcome measures must be adequate to capture the impact of professional development on quality and student gains, and information on reliability must be provided (cf. WWC, 2011). In particular, effect size calculations were found to vary significantly according to the reliability of the scales used (cf. Beelmann, 2014; Hunter & Schmidt, 2004).

Further, *outcome construct* was associated with variance in effect sizes (Wilson & Lipsey, 2001). Even within an outcome domain (e.g., language), the treatment effects were typically much larger for some outcome variables than for others. It is not surprising that treatments have a larger impact on some variables. However, Wilson and Lipsey (2001) show that different outcome constructs within one domain explained 7 % of the variance (p. 418).

The source of information (e.g., teacher or parent ratings, observations, test scores) may confound the effect size calculations. Five % of the variance in effect size estimates were explained when measures were grouped for source of information, including ratings, observations and physical measures (Wilson & Lipsey, 2001). The study by Harvey et al. (2013) investigated predictors of discrepancies between mothers', fathers', and teachers' ratings on behavioral aspects (e.g., hyperactivity, attention problems, aggression) among three-year olds. Mothers', fathers' and teachers' ratings differed significantly and were predicted by ethnicity, ADHD/ODD diagnoses, parental depression, number of children and pre-academic skills.

The *origin of measure*, whether the instrument/scale was developed by the researcher or preexisted as a standardized/published instrument, was tested as a potential effect modifier. Variability due to the origin of the measures was found predominantly in educational research, where researcher-developed measures yielded larger effect sizes (Wilson & Lipsey, 2001). Researcher-developed measures might capture more relevant aspects of changes due to interventions.

It has been suggested that instruments should be sensitive to identify changes and measures should be aligned to interventions to detect effects (Yoon et al., 2007). Indeed, a lack of *alignment between measurement and training* content was associated with smaller effects, at least for caregiver outcomes (Fukkink & Lont, 2007).

Effect size estimation

Further, recommendations on the level of evidence for reviewing studies are made by the WWC (2011; 2013) in order to estimate effects more accurately. First, in quasi-experimental studies, the effect size estimation needs to account for differences in pre-intervention characteristics. Second, for the clear attribution of the effect to the intervention, the intervention cannot be combined with other interventions. Third, the intervention must be implemented as designed (indicating high fidelity) to measure an effect. Forth, the unit of analysis must be considered when examining the impact (cf. WWC, 2011). These recommendations provide methodological issues that can be tested as moderators in the forthcoming meta-analyses.

2.1.5. Integrative framework on professional development effects

The essential need for effective professional development in early childhood education to enhance early childhood educational quality and children's school readiness skills has been acknowledged extensively in the literature (Martinez-Beck & Zaslow, 2006; Sheridan

et al., 2009). Indeed, various professional development programs, differing in respect to focus, form, intensity and other functional features, have been established in early childhood education (e.g., Japel, 2009; Landry et al., 2009; Powell et al., 2010). A shared understanding and consensus is needed “to compare various types of professional development programs or determine which components are most effective, for whom, and under what conditions” (NPDCI, 2008, p. 2). The conceptual framework by Buysse, Winton, and Rous (2009) provides a good foundation for the evaluation of professional development effects and potential moderators. This section targets the integration and systematization of theoretically postulated issues, empirical findings and research needs into an integrated framework (see figure 4) that captures the process from in-service training to child outcomes.

In the integrated framework, the process from in-service training to student achievement consists of multiple steps. First, professional development improves teacher outcomes. *Teacher outcomes* are differentiated between inhibited and visual aspects. In line with adult education theories, teacher knowledge, attitudes, awareness and beliefs, and skills are inhibited outcomes of training (Reischmann, 2004; Weinert, 2001). Teacher performance is “visual” and consists of job performance (e.g., teacher-child interactions, language modeling) that can be measured through observations. It is assumed that attitudes, knowledge and skills predict job performance (cf. Fukkink & Lont, 2007; Yoon et al., 2007).

Second, better job performance enhances the quality of early childhood education. *Quality outcomes* capture educational processes and structures within the classroom. Observational quality measures describe predominately *process quality*, including both physical environmental features and teacher-child interactions (Burchinal, 2010). Based on available instruments, distinctions between general process quality, interaction and domain-specific quality can be made. The quality of early childhood education is mainly investigated through classroom observation methods (e.g., ECERS, ELLCO; CLASS)³ that rate the quality in classrooms available to all children. Further, there are target child observation methods (e.g., ORCE, inCLASS; EAS)⁴ that examine the individual learning experience of a single child in a classroom (ibid.). Both classroom and target child levels are related to children’s development (Howes et al., 2008), but the associations between quality measures and child outcomes are moderate (Burchinal, Kainz, & Cai, 2011). Regarding standards for research, self-reports of quality are not considered as sufficient evidence. In addition to the delineated quality measures that encompass age appropriate activities, the

³ ERS Environmental Rating Scales consist of the ECERS=Early Childhood Environment Rating Scale and the ITTERS=Infant-Toddler Environment Rating Scale; CLASS=Classroom Assessment Scoring System; ELLCO=Early Language and Literacy Classroom Observation.

⁴ ORCE=Observational Record of the Caregiving Environment; inCLASS=Individualized Classroom Assessment Scoring System; EAS=Emerging Academic Snapshot.

proportions of teacher-child interactions and/or quality of instruction and classroom management, *fidelity* has also become prominent as a target outcome of professional development. As mentioned earlier, fidelity is the extent to which target strategy or behaviors are implemented and is predominately linked to effects on student achievement. In the framework, high fidelity is considered as a potential outcome of in-service training and a moderator for training effects on student achievement.

Third, improved quality and high fidelity of intended teaching and interventions lead to enhanced student learning. Expected *child outcomes* of professional development should be aligned with early learning standards. Early learning standards describe learning and teaching goals as well as the learning standards and competencies children should acquire in center-based care before they enter formal schooling (cf. Department for Education, 2012; NAEYC & NAECS/SDE, 2009; OECD, 2012, Scott-Little et al., 2007; Textor, 2012). Early learning standards provide a basis to “promote intellectual, language, physical, social, and emotional development, creating school readiness and building a foundation for later academic and social competence” (NAEYC & NAECS/SDE, 2002, p. 2). In summary, all aspects of child development, health and well-being are potential outcomes for in-service training.

In-service training effects are influenced by professional development components, contextual factors and methodological variables. Starting with key components, the “*who*”-*component* has 2 aspects. One aspect describes who is attending the training and another one focuses on who is facilitating the training. *Trainee characteristics* consist of structural variables (e.g., age, gender, teacher education, experience, ethnicity, and cultural characteristics) and personal variables (e.g., self-efficacy, motivation, knowledge, anxiety, locus of control, cognitive abilities, ability group). Research shows that these features are known to modify intervention effects (cf. Alvarez, Salas, Garofoano, 2004; Blume et al., 2010; Burke & Hutchins, 2007; Cheng & Ho, 2001; Colquitt, LePine, & Noe, 2000; Wilson & Lipsey, 2001). So far, *characteristics of the trainer* (e.g., education, experience, training of trainer, knowledge, theoretical orientations, and perception of role) are assumed to predict in-service education effectiveness (Sheridan et al., 2009, p. 391). However, how these background characteristics influence training effectiveness has not been empirically investigated so far and research is urgently needed in this field. The *relationship aspects* of professional development between trainer and trainee reveal another research gap. At least qualitative research indicates that supportive relationships are critical to intervention success (Brown et al., 2009, p. 501).

The “*what*”-*component* of professional development concerns core competencies (e.g., knowledge, abilities and dispositions) that define training goals, standards and content as well as qualifications (NPDCI, 2008). Training content consists of topics that are needed

to provide high-quality early education to each and every child (NPDCI, 2008; Robert Bosch Stiftung, 2008; 2011; Scott-Little et al., 2007). Recommendations for the general content of in-service training and competence profiles for early childhood educators have been made (Anders, 2012; Buysse, Winton, & Rous, 2009; Fröhlich-Gildhoff, Netwig-Gesemann, & Pietsch, 2011; Robert Bosch Stiftung, 2008; 2011). To ensure the application of new skills in the workplace, professional development content should be coherent with organizational structures and goals as well as the curriculum material that teachers use (Blank & de las Alas, 2009; Buysse, Winton, & Rous, 2009; Elmore, 2002; Garet et al., 2001; Ingvarson et al., 2005; Isner et al., 2011; Yoon et al., 2007; Zaslow et al., 2010a). More specific training content is known to enhance professional development effectiveness (Fukkink & Lont, 2007).

The “*how*”-*component* describes training design features that emphasize aspects of training such as intensity, training principles and delivery format. *Training intensity and duration* have been established as intervention effect modifiers (Wilson & Lipsey, 2001). Results from research into training transfer show that the *training principles* used in professional development are crucial for the application of new skills in the work environment. Learning principles and goals, content relevance, behavior modeling, the provision of post-training interventions and feedback, opportunities to practice and active learning as well as the difficulty level of exercises are known to influence training effectiveness. Other aspects of training principles (e.g., collective participation or manual-based approaches) are assumed to have an influence on training effectiveness, but have not yet been empirically tested. Focusing on *delivery format*, formats can be differentiated into traditional (workshops, courses) and modern approaches (onsite, blended learning, communities of learners). It is assumed that different in-service training methods have differential effects and that each method individually contributes to the impact (Dunst & Trivette, 2011; Varol et al., 2012b); but the evidence base is limited.

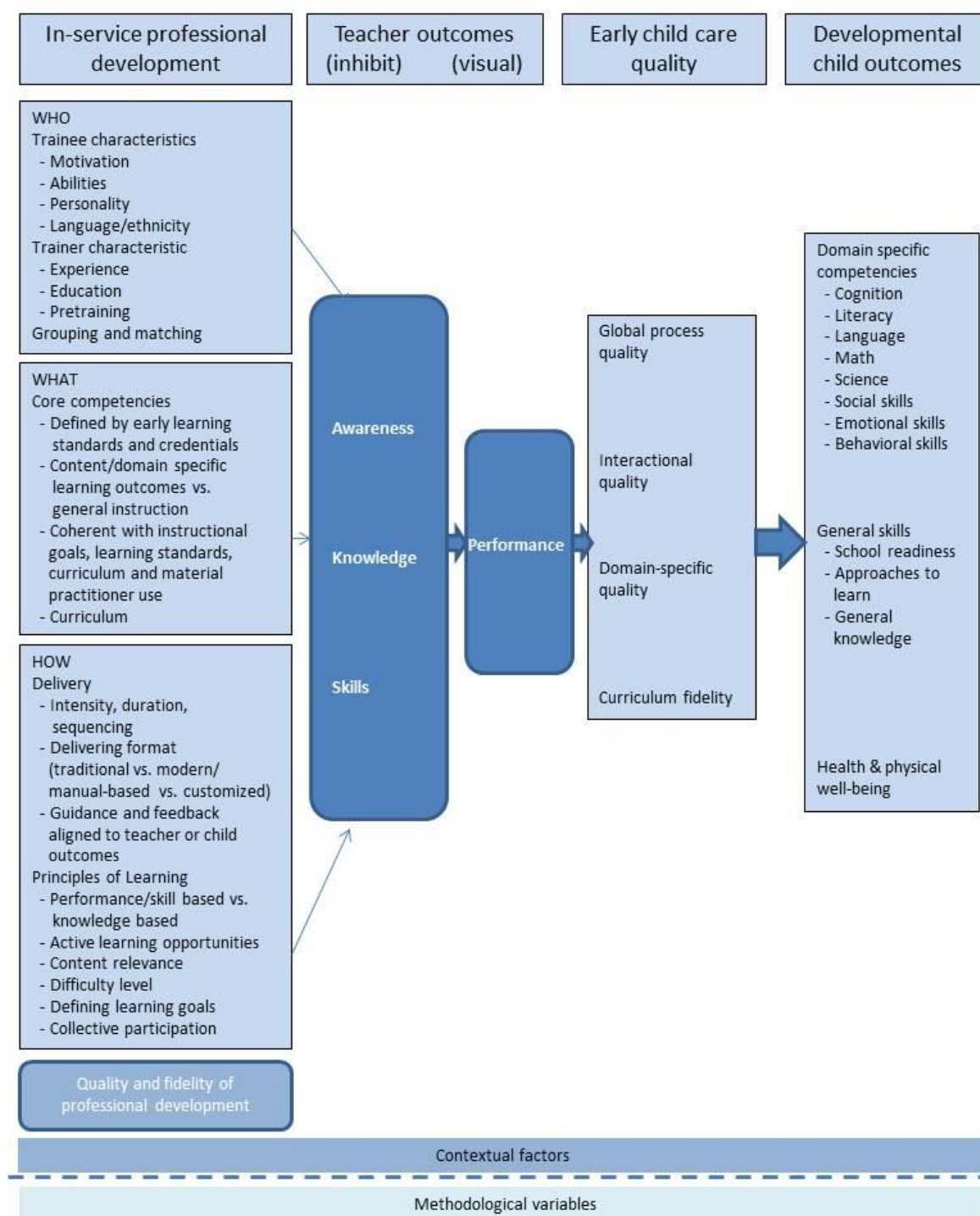


Figure 4. Integrative framework on in-service training effects

The quality and fidelity of professional development are superordinate aspects that strongly influence in-service education effectiveness (Yoon et al., 2007). In line with fidelity research, it is assumed that higher quality professional development and in-service activities that are implemented with higher fidelity carry over into greater effects. While Dane and Schneider (1998) generally suggest that intervention fidelity can be improved through the provision of training manuals, LoCasale-Crouch et al. (2011) demonstrate that ongoing support for trainers and pre-assembled in-service material offer great potential to ensure

the quality of professional development. However, the influence of in-service training fidelity and quality on training effectiveness is widely acknowledged, but the evidence base is limited.

Several *contextual factors* (e.g., resources, policies etc.) are assumed to moderate in-service training effectiveness (Buisse, Winton, & Rous, 2009, NPDCI, 2008). Thus far, organizational structures have been empirically validated as factors modifying training effectiveness in early childhood education. Differential effects were found for different child care systems and providers (Varol et al., 2012a; 2012b) and federal states (Landry et al., 2009), alongside variations in teacher, child, and classroom characteristics. More fine-grained analysis suggests considerable center effects regarding program length and operating hours (Landry et al., 2009). Findings from research into training transfer show that the aspects of a supportive *work environment* and organizational commitment boost training effects. The transfer of new skills is supported when training is consistent with achieving postulated career goals (Burke & Hutchins, 2007; Colquitt, LePin, & Noe, 2000). These contextual variables represent important covariates.

Regarding *methodological issues*, only rigorous study designs can establish the essential foundation to detect in-service education effects (National Research Council, 2004a). Methodological quality influences training effectiveness with regards to design features (e.g., comparison groups, randomization and effect size calculations), measurement features (e.g., informants, reliability, outcome constructs), sample size, and attrition rates (cf. Beelmann, 2014; Wilson & Lipsey, 2001).

To sum up, the *integrative framework* is based on theoretical and empirical findings from training effectiveness research. It offers a solid and comprehensive view of professional development effects in the early childhood education field. Several potential effect modifiers are stressed. Not all potential modifiers have been empirically validated yet. The framework offers a theoretical base for meta-analytic investigations on in-service training effect modifiers. The literature on in-service training impacts on quality ratings and student achievement in the areas of language and literacy as well as social development and problem behavior provides a foundation for empirical evaluation. The following sections review empirical findings from intervention studies on in-service training effects and highlight research objectives to be answered through meta-analysis and meta-regression.

2.2. Review of in-service training effects on quality ratings

A growing demand for early child care has also increased the need for high-quality center-based child care (Autorengruppe Bildungsberichtserstattung, 2012; Laughlin, 2013; Ministère de la Famille et des Aînés, 2007). In particular, a policy focus on academic learning in early childhood has accentuated concerns about poor child care quality. Child care quality is measured by the degree to which early education promotes school readiness or later academic performance. Numerous studies demonstrated that high-quality care predicts positive developmental outcomes for young children (Burchinal et al., 1996; Clarke-Stewart et al., 2002; Howes, 1988; NICHD, 2000a; Peisner-Feinberg et al., 2001). However, the child care and early education received by many children is not of high quality (Karoly et al., 2008; Peisner-Feinberg & Burchinal, 1997; Tietze et al., 2013). In particular, the generally low or at least mediocre quality of child care in western countries (Karoly et al., 2008; Tietze et al., 2013.) has led to calls for efforts to systematically improve quality in early education.

One result of this drive to improve quality has been the development of Quality Rating and Improvement Systems (Zellman & Perlman, 2008). Quality Rating and Improvement Systems (QRIS) combine Quality Rating Systems (quality monitoring and evaluation) with professional development and provide “a sound option for addressing fragmentation by connecting requirements, incentives, and supports that help programs and practitioners improve quality over time” (Connors-Tadros & Carlson, 2011, p. 25). For practitioners, quality ratings are provided along with feedback and technical assistance (Zellman et al. 2008). QRIS typically include the provision of financial incentives or dissemination of information to parents (cf. Boller et al., 2010). Despite their growing popularity, there is little information available on how these systems work or how effective they are (Zellman et al., 2008). Only some pioneer states in the U.S. have evaluated the impact of their Quality Rating and Improvement Systems. The evaluations have shown that quality improves over time, but not all quality changes are statistically significant (Tout, Zaslow, Halle, & Forry, 2009).

The other consequence, beside the state-wide Quality Rating and Improvement Systems, is several regional or national professional development programs. These professional development initiatives can be differentiated between data-driven training and traditional formats not providing data-driven feedback.

Data-driven professional development combines standardized quality ratings with feedback and support based on the data (Bryant et al.; 2009, Zellman et al., 2008). Some of the programs provide data-driven feedback from external quality ratings from common research instruments (e.g., ECERS, ELLCO, or CLASS). Some pilot projects combine a

starting workshop on quality with ongoing data-driven onsite support (Japel, 2009; Manningham, 2008; Palsha & Wesley, 1998). Other programs use adapted versions of quality rating scales for self-assessment as the foundation for ongoing support and guidance (Campbell & Milbourne, 2005; Sheridan, 2001). Self-rating provides a base for reflection and planning for professional development. During the reflection process, teachers share and discuss the ratings (Sheridan, 2001). Teachers identify their individual training needs and determine outcomes that can be targeted through coaching or consultation (Campbell & Milbourne, 2005).

There is a number of traditional non-data-driven professional development approaches. These in-service programs use workshops (Erning, 2003; Japel, 2009), courses (Haskell, 1994) or mentoring (Uttley & Horm, 2008) to improve quality in early childhood classrooms.

The next step is to synthesize findings on the effectiveness of in-service training. A systematic literature search was conducted electronically in data bases like ERIC, PsycINFO, ProQuest Educational Journals, ProQuest Dissertation and Thesis, FIS and WISO, as well as additional hand searches in reference lists, meta-analyses and renowned journals. Overall, 52 studies with 59 different professional development approaches that measured quality through external ratings were found (see table 3). The studies were published between 1992 and 2011 with the majority of studies published in the last 5 years ($n = 33$). Thirty-two of the identified studies were published in journals or books and 20 were dissertations, research reports or conference papers. This indicates a good ratio of 2:3 of published to unpublished studies. Of the 52 studies, 41 studies were (quasi-)experimental studies and 11 studies evaluated the progress of one group. Within the experimental studies, 13 were true RCTs that assigned teachers to treatment or control groups randomly and 11 studies were CRT that assigned classrooms or centers randomly to experimental or control conditions. However, 16 studies compared in-service training conditions with other conditions, but did not use any assignment procedure. Most of the studies evaluated the impact of in-service training with the ELLCO ($n = 21$) or an instrument from the Environmental Rating Scale family ($n = 21$) like the ECERS for preschool and kindergarten classrooms or the ITES for infant-toddler classrooms. In 10 studies, the CLASS was used. The vast majority of studies provided data on the contrast between experimental and control groups at posttest ($n = 32$). Usually inferential statistics (e.g., *t*-test, ANOVA, ANCOVA) were used to calculate the effectiveness of professional development. Only 10 studies reported the extent to which professional development impacts on quality ratings with a standardized effect size and further 11 studies reported descriptive statistics, but did not test for significant differences between the experimental and control groups, nor for pre- and posttest changes.

Variations in professional development strategies

Comparing the 59 professional development treatments, 43 used onsite support (e.g., coaching, consulting, or mentoring) and 14 either a course ($n = 9$) or a single workshop ($n = 5$) to improve early childhood quality. Within the 42 in-service training programs with onsite support, 7 used onsite support as the sole delivery format and 35 used a mixture of onsite support with workshops, courses, online formats or COP. The most common approaches were the combinations of workshop and onsite support ($n = 11$) and course and onsite support ($n = 10$). So far, no professional development approach has tested the impact of solely online training on quality improvements.

Nine studies evaluated programs that focus on the enhancement of general classroom quality with the ECERS/ITERS using onsite support only (Bryant et al., 2009; Uttley & Horm, 2008; Zellman et al., 2008), workshop and onsite support (Japel, 2009; Manningham, 2008; Palsha & Wesley, 1998), course and onsite support (Campbell & Milbourne, 2005; Sheridan, 2001), or the combination of course, workshop and onsite support (Boller et al., 2010). Further, the impact of a workshop on quality was tested by Burchinal et al. (2002) and Erning (2003). An introductory workshop with additional written quality profile without feedback was tested by Japel (2009). Within the 10 quality improvement approaches, 3 used data-driven feedback/onsite support (Boller et al., 2010; Japel, 2009; Manningham, 2008; Zellman et al., 2008), joint needs assessments that facilitated an agenda plan based on a discussion of the results of self-assessment conducted by the teachers and external assessments carried out by the coach (Bryant et al., 2009; Palsha & Wesley, 1998), or self-evaluation on the ECERS scales as the foundation for onsite support (Campbell & Milbourne, 2005; Sheridan, 2001). Several professional development approaches were evaluated ($n = 28$) that emphasized language and literacy. Two thirds used onsite support to help teachers to implement practice designed to improve language and literacy in early childhood classrooms. Seven in-service training programs delivered language-specific contents through a workshop ($n = 1$), courses ($n = 3$), and online modules ($n = 3$).

Further, a great variety of topics and approaches have been developed, accomplished and evaluated to assess the impact of training on quality ratings. These have included classical leadership training, implementations of specific and comprehensive curricula, adaptations of early childhood programs and working with specific populations (e.g., minority boys, first nation children) as well as implementations of behavior management in classrooms. Additionally, unspecific professional development approaches emphasizing issues that cut across early childhood education are widespread. Further, the training effect of a scholarship program (i.e., TEACH) was evaluated, where teachers received financial funding to participate in courses and workshops they were interested in.

Table 3. Review of studies measuring in-service training impacts on quality ratings

Author	Year	Pubtype	WS	CS	ONL	COP	ON	Program	FOC	CUR	Design	Scale
Allgozzine et al.	2011	pub		x			x	Educator PD Partnership Program	CUR	yes	CRT	ELLCO
Barnett et al.	2008	pub	x				x	Tools of Mind Curriculum Implementation	CUR	yes	RCT	CLASS, ERS
Bloom & Sheerer	1992	pub		x			x	Leadership Training Program	LEAD	no	QNE	ECCOS
Boller et al.	2010	unpub	x	x			x	Seed to Success	QI	no	RCT	ERS
Breffni	2011	pub		x				Professional Development Initiative	QI	yes	RCT	CLASS
Brinks	2007	unpub		x			x	Early Reading First/Literacy Instruction	L&L	no	QNE	ELLCO
Bryant et al.	2009	unpub					x	QUINCE - PFI Consultation	QI	no	RCT	ERS
Burchinal et al.	2002	pub	x					Caregiver Training	GEN	no	QNE	ERS
Buysse et al.	2010	pub		x		x	x	Nuestros Ninos PD	L&L	no	CRT	ELLCO
Campbell & Milbourne	2005	pub		x			x	First Beginning	QI	no	QNE	ERS
Cassidy et al.	1995	pub		x				TEACH CDA Program/First Beginning	GEN	no	QNE	ERS
Clancy-Menchetti	2006	unpub	x	x			x	Early Literacy Mentoring	L&L,CUR	yes	CRT	ELLCO
Clancy-Menchetti	2006	unpub						Early Literacy Mentoring	L&L,CUR	yes	CRT	ELLCO
Cunningham (a)	2007	unpub		x			x	TEACH program course & onsite	L&L	no	RCT	ELLCO
Cunningham (b)	2007	unpub		x				TEACH program course	L&L	no	RCT	ELLCO
Dickinson & Caswell	2007	pub		x			x	LEEP Program	L&L	no	QNE	ELLCO
Domitrovich et al.	2009	pub	x				x	Head Start REDI Program	SOC,L&L	yes	CRT	CLASS
Edgar (a)	2008	unpub		x			x	C ³ coaching (University Coaching)	L&L,QI	no	RCT	ELLCO
Edgar (b)	2008	unpub		x			x	C ³ coaching (Master Coaching)	L&L,QI	no	RCT	ELLCO
Englund	2010	unpub					x	Training on Language Modeling	L&L	no	OGPP	CLASS
Erning	2003	unpub	x					Qualitätsentwicklung in Kindergärten	QI	no	OGPP	ERS
Espinosa & Chen	1996	pub		x				Technology & Multiage Grouping	GEN	no	OGPP	PIP
Everston & Smithey	2000	pub	x				x	Mentoring-Protégé Project	SOC	no	QNE	RCI
Fiene	2002	pub	x				x	Infant Caregiver Mentoring Project	QI	no	RCT	ERS
Grace et al.	2008	pub					x	Comprehensive PD on Literacy	L&L	yes	CRT	ELLCO
Hamre et al.	2012	pub		x			x	Support of Language & Literacy	L&L	no	CRT	CLASS
Haskell	1994	unpub		x				All About Child Care	GEN	no	OGPP	ERS

(Table 3. continued)

Author	Year	Pubtype	WS	CS	ONL	COP	ON	Program	FOC	CUR	Design	Scale
Henk et al.	2007	pub			x			HeadsUp! Reading	L&L,CUR	yes	QNE	ERS
Howe et al. (a)	2011	pub					x	Constructivistic curriculum	CUR	yes	QNE	ERS
Howe et al. (b)	2011	pub	x					Constructivistic curriculum	CUR	yes	QNE	ERS
Hsieh	2005	unpub					x	Collaborative Training Model	L&L	no	OGPP	ELLCO
Iutovich et al.	1997	unpub	x	x	x		x	Pennsylvania Child Care Training System	GEN	no	OGPP	ERS
Jackson et al. (a)	2006	pub			x		x	HeadsUp! Reading	L&L	no	QNE	ERS, ELLCO
Jackson et al. (b)	2006	pub			x			HeadsUp! Reading	L&L	no	QNE	ERS, ELLCO
Japel (a)	2009	unpub	x				x	Cap Qualité Pilot (coaching)	QI	no	QNE	ERS
Japel (b)	2009	unpub	x					Cap Qualité Pilot (workshop)	QI	no	QNE	ERS
Justice et al.	2008	pub	x		x			MTP Language & Literacy Curriculum	L&L,CUR	yes	CRT	CLASS
Lonigan et al. (a)	2011	pub	x				x	Mentoring	L&L,CUR	yes	CRT	ELLCO
Lonigan et al. (b)	2011	pub	x					Workshop only	L&L,CUR	yes	CRT	ELLCO
Manningham (a)	2008	unpub	x				x	Cap Qualité Pilot (coaching)	QI	no	CRT	ERS
Manningham (b)	2008	unpub	x					Cap Qualité Pilot (workshop)	QI	no	CRT	ERS
Mason	2010	unpub		x			x	Promoting Academic Success(PAS)	MIN	no	CRT	EAS
McNerney et al.	2006	pub	x				x	Teacher-observation tool guide	L&L	no	QNE	ELLCO
Miller & Bogatova	2009	pub		x				TEACH scholarship program	GEN	no	QNE	ERS
Morris et al	2010	unpub	x	x			x	Foundation of Learning Project	SOC	no	CRT	(in)CLASS
Neuman & Cunningham (a)	2009	pub		x			x	TEACH program course & onsite	L&L	no	RCT	ELLCO
Neuman & Cunningham (b)	2009	pub		x				TEACH program course	L&L	no	RCT	ELLCO
Neuman & Wright (a)	2010	unpub					x	TEACH program course & onsite	L&L	no	RCT	ELLCO
Neuman & Wright (b)	2010	unpub		x				TEACH program course	L&L	no	RCT	ELLCO
Onchwari & Keengwe	2010	pub					x	Strategic Teacher Education Program	L&L	no	QNE	ELLCO
Palsha & Wesley	1998	pub	x				x	Consultation model for inclusion	QI,	no	OGPP	ERS
Pianta et al.	2008	pub			x		x	My Teaching Partner (MTP)	L&L,SOC	yes	RCT	CLASS
Powell et al. (a)	2010	pub		x	x		x	Classrooms Linked to Early Literacy	L&L	no	RCT	ELLCO
Powell et al. (b)	2010	pub		x	x		x	Classrooms Linked to Early Literacy	L&L	no	RCT	ELLCO

(Table 3 continued)

Author	Year	Pubtype	WS	CS	ONL	COP	ON	Program	FOC	CUR	Design	Scale
Powell et al.	2009	unpub	x				x	Classrooms Linked to Early Literacy	L&L	no	RCT	ELLCO
Raver et al.	2008	pub		x			x	Chicago School Readiness Program	SOC	no	RCT	CLASS, ERS
Roehrig et al.	2011	pub		x		x	x	Ah Neen Dush	SCI, QI	no	OGPP	CLASS
Sheridan	2001	pub		x			x	Model of Competence Development	QI	no	QNE	ERS
Sibley & Sewell	2011a	pub	x	x		x	x	Multidimensional PD	CUR	yes	QNE	ELLCO
Sipp	2010	unpub		x			x	Emotional & Instructional Support	SOC, QI	no	OGPP	CLASS
Sprague	2009	unpub		x			x	Community-Advocacy PD	CUR	yes	CRT	ELLCO
Uttley & Horm	2008	pub					x	RICDSAP	QI	no	QNE	ERS
Wasik & Hindman	2011	pub	x	x			x	ExCELL Coaching Program	L&L	no	CRT	ELLCO, CLASS
Zellman et al.	2008	unpub					x	QUALISTAR	QI	no	OGPP	ERS, Snapshot

Note. Pubtype=publication type; pub=published in journals or as book chapter; unpub=unpublished paper e.g., technical paper, dissertation or governmental report; WS=workshop; CS=course; ONL=online training or blended learning; COP=community of learning or community of practice; ON=onsite support; FOC=focus of training; CUR=curriculum implementation; INS=quality rating instrument; L&L=language and literacy; SOC=social and behavioral development; QI=quality improvement initiative; SCI=science; GEN=general focus of early childhood education(e.g., developmentally appropriate practice); RCT=randomized controlled trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center to groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design ; OGPP=one group pre-post measure; ERS Environmental Rating Scales (e.g., ECERS or ITES); ELLCO=Early Language and Literacy Classroom Observation; CLASS=Classroom Assessment Scoring System; (in)CLASS=Individualized Classroom Assessment Scoring System ECCOS=Early Childhood Classroom Observation Scale (Bredekamp, 1986); PIP=adapted Version of the Primary Implementation Profile (PIP) Form (High/Scope, 1989); RCI=Rating of Classroom Instruction (Everston & Burry, 1989); EAS=Emerging Academic Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2001); Snapshot=Pre-Kindergarten Snapshot (Howes, 1997).

2.2.1. In-service training impacts on process quality ratings with Environment Rating Scales

Environment Rating Scales (ERS) assess process quality in early childhood classrooms primarily through observations. Process quality consists of various interactions among children, staff, parents and material provided in the classroom. Features of the learning environment in early childhood classrooms (e.g., material, space, activities and daily schedules) are assumed to support these interactions. Process quality is found to be a strong predictor for developmental outcomes in young children (NICHD, 2000a; Peisner-Feinberg et al., 2001). So far, 21 studies measured the impact of in-service programs on process quality in infant-toddler, preschool and kindergarten classrooms with Environment Rating Scales.

Impact of in-service programs on process quality in infant-toddler classrooms

The impact of in-service programs on the quality of infant-toddler classrooms has been evaluated by 7 studies. Only 3 studies provided detailed information on changes in total scores and subscales. However, the findings within the studies on total and sub-scores were inconsistent. Campbell and Milbourne (2005) reported a significant effect for time and consultation conditions on the ITERS total score for the First Beginning Program (course and consultation). However, this interaction effect might be due to the fact that in the control group most of the ITERS scores decreased, while the group in the consultation condition showed only slightly increased results (ibid.). Fiene (2002) did not find a significant impact on the ITERS total score, but the subscales “personal care routines” and “activities” changed over time through an initial workshop and continuing mentoring. The impact of a consultation model on ITERS scores in child care centers with inclusion was examined by Palsha and Wesley (1998). The teachers participated some initial training on how to administrate the ITERS. Furhter, teacher received 12 consultation sessions over a period of 6 to 12 months as well as 200 USD financial support. Significant effects were substantiated for the ITERS total score and the subscales “activities”, “space and furnishing”, and “listening and talking”. Additionally, the impact of the Early Childhood Development Training System in Pennsylvania was evaluated by quality ratings of infant-toddler classrooms. Iutcovich et al. (1997) reported that the quality of infant-toddler classrooms was not predicted by the amount of training teachers received (state regulation of 6 hours per year), but was predicted by higher salaries. However, teachers participating in courses at the community college with approximately 9 to 11 credit hours showed significant gains in the ITERS subscales “interaction” and “space & furnishing”, but not on the total score (Miller & Bogatova, 2009). Several child care centers run services for children from birth to the age of 5 or 6. In 3 studies, a combined score of ITERS-R and ECERS-R was used to measure the quality of child care centers. The results by Burchi-

nal et al. (2002), comparing 553 classrooms in center-based care, clearly indicate that teachers who attended workshops at the center, the community or at professional meetings had higher ITERS-R and ECERS-R total scores. The Early Childhood Associate Degree Scholarship Program, a program designed for employees in early child care who wish to pursue a CDA degree, was evaluated using the Environmental Rating Scales. Teachers were required to complete 12 to 20 credit hours at the community college (Cassidy et al., 1995). Teachers in the scholarship program obtained significantly higher scores on ITERS-R and ECERS-R total scores and the “developmentally appropriate activities” subdomain, but not on the “appropriate caregiving” scale. Further, the Seeds to Success QRIS was piloted in a 7-month period in Washington State. The QRIS was composed of 8 hours of coaching per month, quality improvement grants based on the quality ratings (1,800 to 12,600 USD per center) and funds for professional development opportunities as well as release time and expenses. Significant impacts were found in the combined ITERS-R and ECERS-R total score as well as all subscales on the Environment Rating Scales (Boller et al., 2010). An exploratory analysis indicated that the impact on the ITERS subscales seemed to be driven at item level, and in particular, large significant changes in the subscale “activities” were driven by the items “dramatic play”, “nature and science”, and “diversity”. This information might be useful for further individual professional development programs such as coaching or technical assistance.

Impact of in-service programs on the quality of preschool and kindergarten classrooms

As mentioned above, inconsistencies occurred within studies and subscales. The improvement of preschool and kindergarten process quality through attendance at *workshops* was examined in Canada, Germany and the United States. Burchinal et al. (2002) demonstrate that attendance at any workshop at centers, communities or professional meetings makes a difference to quality ratings on the ECERS-R total score in the United States. Descriptive findings by Erning (2003) also indicate that quality ratings on the ECERS-R total and subscales in center-based early childhood classrooms in Germany can be improved through a series of workshops based on the principles of the ECERS, delivered over a period of 2 years. Teachers at 3 different sites in Canada received 4 2-hour workshops to implement a constructivist curriculum (Howe et al., 2011). Teachers from the workshop condition did not reach significantly higher scores on the ECERS-R subscales “interaction” or “activities” than the control or onsite conditions at posttest, though the ECERS-R scores at pretest were considerably lower for the workshop condition (ibid.). The findings from the Canadian Cap Qualité pilot project provided evidence that a) the ECERS-R total score could be improved through a workshop on quality for teachers and administrators and b) the effects could be maintained a year after the initial training in contrast to the control group (Japel, 2009; Manningham,

2008). The descriptive findings indicated that higher gains could be achieved through the combination of workshop and additional onsite support (ibid.).

Studies that analyzed the impact of a *workshop and additional onsite support* on ECERS-R ratings showed significant changes on the ECERS-R total score through in-service training. However, statistically significant effects were not achieved for all subscales. Barnett et al. (2008) found significant differences on the ECERS-R total score as well as some subscales (e.g., “activities”, “interactions”, and “language and reasoning”) for a 4-day curriculum training program (Tools of Mind) and weekly 30-minute classroom visits throughout the school year. Statistically meaningful contrasts at posttest on the ECERS-R total score were also reported for a workshop on quality for child care teachers and administrators and 3 data-driven sessions of coaching (Japel, 2009; Manningham, 2008). Only the consultation approach by Palsha and Wesley (1998), including a training session on the ECERS-R and data-driven coaching, gained significant changes on the ECERS total score and all subscales in community-based centers that aim to provide inclusive services.

Several onsite support models like consultation or technical assistance have been used to improve process quality. Teachers in the Quality Intervention for Early Care and Education QUINCE study received assessment-based consultations over a period of 6 months (Bryant et al., 2009). At posttest, there was no significant contrast between the consultation and control conditions. However, significant changes in the ECERS-R factor score “provision for learning” were found for pretest and posttest gains as well as follow-up gains (ibid.). In addition to the ECERS-R, the ECERS-E was assessed. At posttest, a significant contrast could be seen for the ECERS-E total score and subscale “literacy” ($d = 0.34$), but not for “numeration” (Bryant et al., 2009, p. 67). Another consultation model was tested by Howe et al. (2011). During the 15-week intervention, teachers received a training manual and approximately 2 to 4 hours of individual support, including classroom observation, feedback and modeling to implement a constructivist curriculum. Significant contrasts or changes over time were not found for the consultation condition (Howe et al., 2011). The impact of the Rhode Island Child Development Specialist and Apprenticeship Program (RICDSAP), using an apprentice-mentor model, showed gains on the ECERS-R total score after 6 months of mentoring (Uttley & Horn, 2008). In the QUALISTER QRIS in Colorado, teachers received technical assistance to improve quality sustainably based on their ratings (Zellman, 2008). However, the findings reported are ineligible for inclusion, because of inappropriate test scheduling with variations of up to 20 weeks in the 12-month testing intervals (ibid.).

Three studies have confirmed the impact of *courses* on quality ratings on the ECERS-R total score. Descriptive data suggests that quality on the ECERS-R total and subscales can be improved through a course with 9 4-hour long sessions provided over a period of 10 weeks (Haskell, 1994). Cassidy et al. (1995) found significant gains on the ECERS-R as well as the

ITERS-R total scores and the subdomains “appropriate caregiving” and “developmentally appropriate activity” for teachers who participated in the Early Childhood Associate Degree Scholarship program (see T.E.A.C.H.) and attended 12- to 20-credit-hours-courses at community colleges. Teachers in the T.E.A.C.H. Early Childhood ® Project Pennsylvania who attended courses at community colleges with approximately 9 to 11 credit hours showed significant changes over time on the ECERS-R total and all subscales, except “personal care routines” (Miller & Bogatova, 2009).

The impact of *online and satellite-based training* in the HeadsUp! Reading (HUR) program on the quality ratings of preschool and kindergarten classrooms was assessed in Missouri and Nebraska. HUR is a satellite-based distance-learning instructional course for early childhood teachers focusing on research-based strategies to support children’s literacy development free of charge. Teachers in Nebraska reached significantly higher scores on the ECERS total score and the subscales “personal care”, “activities”, and “parent and staff” compared to a control group (Jackson et al., 2006). However, teachers participating in the HUR-training sessions in Missouri did not show meaningful changes on the ECERS-R total score (Henk et al., 2007). In Nebraska, a self-selected group of teachers (n = 17) received mentoring in addition to the HUR-course. The HUR-mentoring group did not improve their ECERS-R scores significantly more than the control group (Jackson et al., 2006).

Additionally, an intensive professional development approach, called Model of Competence Development, was evaluated in Sweden by Sheridan (2001). Teachers received a) a course of 8 monthly lectures, b) support in facilitating a learning organization (COP), and c) individual guidance based on discussions of self- and external evaluations on the ECERS (ibid.). The descriptive data indicate that quality can be improved over time through this intensive approach.

2.2.2. In-service training impacts on language and literacy-specific quality

Overall, 21 studies with 27 different approaches evaluated the impact of professional development on language and literacy-specific quality, measured with the ELLCO. Of these, 22 in-service training formats used individual onsite support (e.g., coaching, consultation, mentoring) to help teachers to increase language and literacy practice in early childhood classrooms. Five in-service programs did not rely on individual support and used courses (n = 3), workshops (n = 1) as well as online formats (n = 1). The findings point out that it is possible to change language and literacy-specific quality sustainably in ECCE classrooms (Brinks, 2007; Cunningham, 2007; Dickinson & Caswell, 2007; Hsieh, 2005; Wasik & Hindman, 2011) and certain features like literacy activities (Brinks, 2007; Buysse, Castro, & Peisner-Feinberg,

2010; Dickinson & Caswell, 2007; Hsieh, 2005; Jackson et al., 2006; Onchwari & Keengwe, 2010), literacy material in classrooms (Brinks, 2007; Dickinson & Caswell, 2007; Jackson et al., 2006; Neuman & Wright, 2010; Sibley & Sewell, 2011a; 2011b; Sprague et al., 2009) and the classroom environment (Brinks, 2007; Dickinson & Caswell, 2007; Grace et al., 2008; Lonigan et al., 2011; Sibley & Sewell, 2011a) through professional development. However, some professional development formats seem to be more effective than others.

The effectiveness of a 45-hour, 3-credit-bearing course on language and literacy (Cunningham, 2007; Neuman & Cunningham, 2009) as well as an adapted 3-credit course of 30 hours (Neuman & Wright, 2010) was investigated. After controlling for pretest values, in both studies the ELLCO scores at posttest for the group receiving a course remained statistically equivalent to the scores of the control group. So far, there is no evidence for the impact of *credit-bearing courses* at a local community college on language-specific quality in early childhood classrooms. However, the findings from Lonigan et al. (2011) suggest that small changes in several ELLCO scores could be achieved through the delivery of curriculum material (Literacy Express Preschool Curriculum; LEPC) and curriculum-specific professional development consisting of a 2-day *workshop* prior to the school year and a half-day workshop during the school year. Further, teachers participating in a 15-week satellite broadcast HeadsUp! Reading (HUR) *online/blended learning* course improved significantly more on the ELLCO literacy environment checklist than the control group, but not on other ELLCO scores (Jackson et al., 2006). The findings from studies investigating professional development formats without onsite support showed that only marginal improvements in language-specific quality can be achieved. The results also suggested that individual support based on classroom needs is required to produce changes in classroom quality.

Four studies investigated the impact of onsite support as main strategy. A 10-week coaching approach with 3 hours of coaching a week showed a significant impact on some items on the ELLCO (e.g., books, writing material) at posttest. Enlarged effects could be seen at follow-up for the ELLCO subscale “books” (Neuman & Wright, 2010). In the pilot study from Hsieh (2005), 5 teachers received individual support with observational feedback to implement 3 different skill clusters (A: vocabulary and narratives; B: phonological awareness and alphabet; C: print concept and written language) to promote language and literacy development. At the end of the study, the teachers demonstrated the learned skills with relatively high fidelity (> 75 %) and changes in practice led to higher scores on the ELLCO total score as well as the classroom observation and literacy activity rating subscales, but not on the literacy environment checklist (ibid.). In addition, a huge impact of weekly data-driven coaching activities (approximately 5 hours) plus curriculum material and financial funds on the ELLCO total score and all subscales was found by Grace et al. (2008). However, onsite support activities provided twice a month for approximately 4 hours by a mentor-coach only led to a significant

contrast on the ELLCO General Classroom Observation scale, but not on other the ELLCO subscales (Onchwari & Keengwe, 2010). These findings suggest that onsite support alone can change literacy practice, but further in-service efforts (e.g., curricula, material, funds, or other delivery formats) are needed to gain major changes.

Approaches providing *courses and additional onsite support* seem to be very successful in changing language-specific quality. Significant effects for combined approaches, including a 45-hour 3-credit-bearing language and literacy course at a community college and additional coaching (2 hours weekly for 32 weeks), were reported on the ELLCO total score (Cunningham, 2007) and several ELLCO items (Neuman & Cunningham, 2009) in contrast to control groups at posttest. Large significant impacts on all ELLCO scores were achieved by teachers participating in an intensive professional development program that provided monthly meetings and weekly coaching activities over a period of 3 school years (Brinks, 2007). Further, teachers that participated in the HeadsUp! Reading (HUR) course, the Opening the World of Learning (OWL) course (approximately 75 hours), and a 30-hour mentoring program scored statistically considerably higher on the ELLCO subtest "Literacy Environment Checklist" and the ELLCO subscale "Language, Literacy, Curriculum", but not on other subscales or the total score (Sprague et al., 2009). Teachers in the Early Childhood Educator Professional Development (ECEPD) Project received the Open the World of Learning (OWL) curriculum and continuous, intensive and extensive in-service training through large and small group sessions and coaching (Algozzine et al., 2011). Due to methodological issues concerning a treatment group with very high means at pretest in comparison to a control group ($d = 0.90$), no significant effects were found for changes in ELLCO scores over time, but there was an expected contrast ($d = 0.48$) seen at posttest (ibid., p. 253). On the other hand, Edgar (2008) demonstrated that significant changes and contrasts on several ELLCO items and subscales were gained through the 9-week University Coach Approach that offers a course with bi-monthly meetings and individual coaching sessions. Further, the study by Edgar (2008) provided empirical evidence on the importance of the qualification level of the coach. While significant results on several ELLCO items were achieved when university and research staff did the training (University Coaching Approach), non-significant changes were gathered through a multiplier training format, where the participants of the University Coaching Approach acted as coaches in the following year (Master Coaching Approach).

Findings from a short-term approach, using a 2-day *workshop* and semester-long *onsite activities* (approximately 7 coaching sessions), demonstrated significantly large increases on 2 ELLCO subscales "Literacy Environment Checklist" and "General Classroom Observation" (Powell, Diamond, & Burchinal, 2009; Powell, Diamond, Burchinal, & Koehler, 2010). The assumption of the effectiveness of in-service programs combining workshops and onsite support was confirmed by Lonigan et al. (2011). They demonstrated the impact of an initial 2-

day workshop plus half-day booster workshop and school year-long mentoring and additional curriculum material on all ELLCO scores ($d = 0.14$ to 0.44) in comparison to a control group. However, on some ELLCO scores the workshop group made greater gains than the teachers in the workshop plus mentoring condition (ibid.). Dickinson and Caswell (2007) investigated the impact of a 45-hour, 4-credit-bearing course delivered through 2 workshops that were 3 days long and supervision as well as 4 performance-based assignments for teachers to integrate their new understanding into daily classroom practice. The teachers in the workshop and supervision approach achieved higher values in all ELLCO scores. McNerney et al. (2006) used a differential analytic approach to investigate improvements and diversity between schools participating in a year-long in-service program with a series of workshops and weekly coaching activities guided by literacy coaches. The descriptive findings clearly indicated great diversity in change rates on ELLCO scores between schools and teachers within a school (ibid., p. 24). McNerney and colleagues (2006) suggested that ELLCO scores could be used in addition to data-driven coaching in a formative way a) for systematic material purchasing in classrooms and b) for the adjustment of professional development workshop sequencing and delivery. A *remote coaching* approach (distance coaching with teacher videos and online feedback) in addition to the 2-day *workshop* was also tested by Powell et al. (2010). The results indicated meaningful gains for the remote condition on one ELLCO subscale, but in comparison to onsite coaching, teachers in the remote group made less progress ($d = 0.11$ to 0.23). Considering the variety of onsite formats, different coaching models or onsite support approaches (mentoring vs. coaching) might influence the effect size sustainably and should be further analyzed.

Multidimensional professional development strategies were used in 3 studies. In the PEACH Partnership Program, teachers received an in-depth workshop, coursework, individual onsite support, and a learning group to implement the OWL-curriculum and a literacy and numeracy-enriched environment and to integrate child assessment for individual curriculum planning (Sibley & Sewell, 2011a). Teachers from the treatment group scored higher on the ELLCO than the comparison group, but the contrast was not significant (ibid.). To implement the Nuestros Ninos model to foster Dual Language Learners in early childhood classrooms, teachers participated in an initial workshop and continuous community of practice meetings and obtained ongoing consultation. Significant effects occurred on the ELLCO Literacy Activity subscale and some items on an ELLCO Addendum, but not for other literacy environment or general literacy practice items (ibid., p. 200). The Exceptional Coaching for Language and Literacy (ExCELL) professional development model consists of 4 half-day institute and ongoing training cycles, including group training, in-class behavior modeling by the coach and in-class observation and feedback and distinctively affects language and literacy-specific quality (Wasik & Hindman, 2011). Teachers from the 15-week (HUR) satellite based (online) course

that additionally received mentoring (6 times, 2 to 4 hours long over 2 months) scored significantly higher on the ELLCO “Literacy Activity Ratings Scale” and the “Literacy Environment Checklist” (Jackson et al., 2006). Clancy-Menchetti (2006) compared 2 types of intensive in-service program that included an initial workshop and ongoing meetings and one group received additional mentoring. The results indicated no significant differences between the intervention conditions, but Head Start classrooms scored lower than public classrooms.

2.2.3. In-service training impacts on interaction quality ratings with the CLASS

So far, 12 studies have evaluated the impact of different professional development approaches with a focus on language and interactions using the Classroom Assessment Scoring System (CLASS). All of them used multidimensional professional development with on-site support, except Breffni (2011) who evaluated an 8-week course. However, not all studies investigated the effectiveness of in-service training on the full range of domains and subscales of the CLASS. Therefore, the results are arranged by quality domain (see table 4).

The findings from the evaluations using the CLASS are inconsistent. Eight studies evaluated the impact of professional development on Emotional Support in classrooms, but only 5 studies found significant effects. Five out of 7 studies reported significant changes in the categories of Classroom Organization and Instructional Support through in-service programs.

The dimension *Emotional Support* includes the items “positive classroom climate”, “negative classroom climate”, “teacher sensitivity” and teachers’ “regard for student perspective”. Emotional Support is associated with higher values in social competencies and letter naming and reduced problem behavior (Curby et al., 2009; Dominguez et al., 2011; Downer et al., 2011b). Empirical findings demonstrated that Emotional Support can be enhanced through courses, either 8 or 14 weeks long (Breffni, 2011; Hamre et al., 2012), but the effect was not seen on every item (Hamre et al. 2012). The Foundation of Learning (FOL) Project, using the Incredible Years Program (see Webster-Stratton, 2000) and clinical consultation, increased teachers’ sensitivity and decreased negative climates considerably (Morris et al., 2010). Further, the My Teaching Partner (MTP) consultation approach resulted in significant improvements in teacher sensitivity (Pianta et al., 2008). Teachers in the Chicago School Readiness Project showed higher levels of positive classroom climate and lower levels of negative climate than controls (Raver et al., 2008). Teachers who participated in the Ah Neen Dush professional development program for centers in the American Indian Reservation for 2 years scored significantly higher on the CLASS subscale “positive climate” and “regard for student perspective” (Roehrig et al., 2011). Implementing the Tools of Mind curriculum (Barnett et al.,

2008), the REDI-program (Domitrovich et al., 2009) or the EIS-program (Sipp, 2010) did not show significant changes in emotionally supportive teaching practice.

Table 4: Evidence base of in-service effects on CLASS ratings

Author	Year	ES	PC	NC	TS	RSP	CO	BM	PD	ILF	IS	CD	QF	LM
Barnett et al.	2008		n.s.	n.s.	n.s.									
Breffni	2011	+					+				+			
Domitrovich et al.	2009	n.s.							n.s.	n.s.	n.s.		n.s.	
Englund	2010													+
Hamre et al.	2012	+	n.s.	n.s.	n.s.	+	n.s.	n.s.	n.s.	+	+	+	+	+
Morris et al.	2010		n.s.	+	+	n.s.		n.s.		+			n.s.	n.s.
Pianta et al.	2008		n.s.	n.s.	+	n.s.		n.s.	n.s.	+		n.s.	n.s.	+
Raver et al.	2008		+	+	n.s.			+						
Roehrig et al.	2011	n.s.	+	n.s.	n.s.	+	+	n.s.	+	+	+	n.s.	+	n.s.
Sipp	2010		n.s.	n.s.	n.s.	n.s.		n.s.	n.s.	n.s.		n.s.	n.s.	n.s.
Wasik & Hindman	2011										+	+	+	+

Note. +=significant contrast or change score; n.s.=not significant; *=significant after 2 years of training; **=significant after 3 years of training ES=emotional support; PC=positive climate; NC=negative climate; TS=teacher sensitivity; RSP=regard for student perspective; CO=classroom organization; BM=behavior management; PD=productivity; ILF=instructional learning format; CD=concept development; QF=quality of feedback; LM=language modeling.

The dimension *Classroom Organization* consists of 3 items to rate the quality of “behavior management”, “productivity”, and “instructional learning formats” in classrooms. Classroom Organization is strongly related to social competencies, math skills, letter naming, emergent language and literacy development, and lower values of problem behavior. Further, children with or without an immigration background benefit from well-organized classrooms (cf. Dobbs-Oates et al., 2011; Downer et al., 2011a). Enhanced Classroom Organization was targeted by 8 in-service approaches. However, only 2 approaches were effective (Breffni, 2011, Raver et al., 2008). A small amount of studies demonstrated some changes (Hamre et al., 2012; Morris et al., 2010; Pianta et al. 2008; Roehrig et al., 2011) and 2 training formats produced no effect on Classroom Organization (Domitrovich et al., 2009; Sipp, 2010). An 8-week course on “Best Practice in Prekindergarten Curriculum” (Breffni, 2011) showed significant effects on Classroom Organization. The Chicago School Readiness Program that implemented a behavior management system with simple coursework and coaching (Raver et al., 2008) improved the CLASS “behavior management” subscale appreciably. No statistically significant impacts were found for the REDI-program (Domitrovich et al., 2009) and the EIS-approach (Sipp, 2010). Teachers in the Ah Neen Dush in-service program, organized as a series of monthly meetings, made significant changes from year one to year 2 in Classroom Organization, enhanced their productivity and instructional learning formats, but did not change behavior management in the classroom (Roehrig et al., 2011). A significant increase in child-centered instructional learning formats in classrooms was confirmed by 3 training programs. Hamre et al. (2012) found a considerable effect with a 14-week course and year-

long consultation and Morris et al. (2010) did the same with the Foundation of Learning (FOL) approach, including an Incredible Years course, stress management workshop, and clinical consultation. Pianta et al. (2008) also found an effect with the MTP online consultation condition.

The dimension *Instructional Support* consists of quality ratings on “concept development”, “quality of feedback”, and “language modeling”. High values of Instructional Support predict receptive vocabulary, oral and written language, rhyming, letter naming and applied problem solving skills (Mashburn et al., 2008). In particular, higher values in “concept development” are strongly related to receptive vocabulary and applied problem-solving skills (Curby et al., 2009). Nine studies evaluated the impact of in-service programs on Instructional Support or single items. Of these, 4 studies demonstrated the effectiveness of professional development (Breffni, 2011; Englund, 2010; Hamre et al., 2012; Wasik & Hindman, 2011), 2 approaches found some evidence (Pianta et al., 2008; Roehrig et al., 2011), and 3 training formats did not result in significant improvements (Domitrovich et al., 2009; Morris et al., 2010; Sipp, 2010). The Exceptional Coaching for Early Language and Literacy (ExCELL) program (Wasik & Hindman, 2011), with summer institutes, ongoing training cycles and coaching as well as the professional development 14-week course and year-long consultation program (Hamre et al., 2012), showed significant improvements in the Instructional Support domain and all subscales. Additionally, the 8-week best practice course (Breffni, 2011) tested and verified the impact of the Instructional Support domain. A single-subject design was used to test the impact of an individual coaching approach with 4 sessions on language modeling (Englund, 2010). The descriptive findings demonstrate a considerable increase on the performance of language modeling for all teachers. Follow-up measures clearly indicate that the teachers were able to maintain their performance levels (ibid., p. 146). Teachers in the MTP consultation condition achieved significant improvements on “language modeling” scores, but not on “concept development” or “quality of feedback”. To receive considerable effects on Instructional Support and the subscale “quality of feedback”, teachers had to participate in a third year in the Ah Neen Dush program (Roehrig et al., 2011). Nevertheless, 3 years of intervention did not change teacher practice in “concept development” or “language modeling” (ibid.). The REDI program (Domitrovich et al., 2009), the Emotional and Instructional Support (EIS) approach (Sipp, 2010), and the FOL training program (Morris et al., 2010) did not result in any significant changes in any aspect of Instructional Support.

To sum up, there are inconsistencies within and between studies. One potential explanation is the alignment (or not) of treatments to the CLASS scores used as the preferred outcome. Fukkink and Lont (2007) point out that a close match between treatments and measures is related to the magnitude of in-service training effects. Another reason might be fidelity to the

professional development program. In particular, a high degree of in-service training fidelity was related to differences (improvements) in trainees' use of target strategies and instructional practices (Trivette, Raab, & Dunst, 2012a; 2012b). When trainers used the recommended coaching practices of the PALS-Program to a high degree, teachers had significantly higher scores ($d = 0.32$ to 0.67) in Emotional Support, Classroom Organization, and Instructional Support in contrast to a low fidelity group (ibid., p. 54)

2.2.4. In-service training impacts on alternative classroom quality ratings

Three studies measured the impact of professional development programs with other instruments than the CLASS, the ELLCO, or the Environmental Rating Scales. Bloom and Sheerer (1992) used a modified version of the Early Childhood Classroom Observation Scale *ECCOS* (Bredekamp, 1986) to assess the impact of the Early Childhood Leadership Training Program for teachers and directors on the quality of teaching practices. Teachers who took part in the intervention made significantly more progress than a control group on the total score and all subscales (interactions among staff and children, curriculum, health and nutrition, physical environment). The effectiveness of the Constructing and Networking for Multi-age Learning training program for elementary school teachers (kindergarten to 6th grade), including monthly workshops on multi-age grouping and a technology workshop, was measured with an adapted version of the Primary Implementation Profile (*PIP*) Form (High/Scope, 1989). Teachers significantly improved the quality of daily routines, adult-child interactions and overall classroom quality, but the quality of the physical environment did not change considerably (Espinosa & Chen, 1996). A guided mentoring approach was used by Everston and Smithey (2000) to improve novel teachers' management practices, lesson planning, and goal-setting skills. The effect was investigated with an observation tool called Rating of Classroom Instruction *RCI* (Everston & Burry, 1989). The *RCI* measures the arrangement of the physical setting, management of instruction, establishment of rules and procedures, motivating and managing student behavior as well as classroom climate. The teachers, who took part in the intervention, teaching in kindergarten through high school, reached significantly higher scores at posttest than the control-group teachers (ibid., p. 300).

2.2.5. In-service training impacts on target-child quality

Target-child quality, in particular the context of the child's interactions with teachers and peers, describes the daily classroom experience of a single child that contributes to social and academic outcomes. Only 2 studies have measured the impact of in-service programs on quality ratings with target-child-observations. Morris et al. (2010) used the Individualized Classroom Assessment Scoring System (inCLASS) by Downer et al. (2010) that measures quality at target-child level. The inCLASS assesses child behavior in interactions with the teacher. The Children in Foundation of Learning (FOL) training condition had significantly lower rates of teacher and peer conflict than children in control classrooms. Further, the children's task engagement and task behavior control were considerably higher in FOL classrooms. However, no significant contrast could be reported for peer-related interactions (communication, sociability, and assertiveness), teacher communication and positive engagement as well as children's task self-reliance (ibid., p. 36).

Twenty-three pre-kindergarten to third grade teachers participated in the 5 training days of the "Professional Development Promoting Academic Success for Boys of Color (PAS)" approach (Mason, 2010). The intervention effect was measured with the Emergent Academic Snapshot EAS (Ritchie, Howes, Kraft-Sayre, & Weiser, 2001). The intervention group had significantly higher scores on the "oral language development" and "attentive" scales, but not on "scaffolding". PAS training explained between 27 – 30 % of the variance in the target child's scores on the EAS (Mason, 2010, pp. 68-69).

2.2.6. Challenges measuring in-service training effects on quality ratings

The review of intervention studies showed that the findings on the impact of professional development on quality ratings are inconsistent. Inconsistencies were found for outcome constructs (e.g., instrument or score level). Differences in effect sizes due to outcome constructs are not surprising, because some outcomes are more closely aligned with treatments and larger impacts would be expected. However, it is not possible to estimate how outcome constructs influence the magnitude of in-service training effects.

Further, clear conclusions cannot be drawn at this stage on the most effective professional development model to improve quality. Findings from the ratings on the ERS indicate that effectiveness is moderated by onsite support (e.g., mentoring, coaching, or consultation). In the literature, coaching and other individualized onsite support methods (e.g., consultation, mentoring, technical assistance) are considered as the most promising professional development approaches to enhance overall quality in early childhood settings and the application

of new teaching practice (Powell & Diamond, 2011, Powell, Diamond, & Burchinal, 2009; Sheridan et al., 2009; Tout, Isner, & Zaslow, 2011). While most of the time, coaching and consulting are used interchangeably in the early childhood literature (Forry, Tout, Zaslow, & Martinez-Beck, 2011; Llyod & Modlin, 2012), coaching can be differentiated from other individualized onsite approaches through its specialized support using a series of activities, including joint planning, observation, feedback, and reflection (Gupta & Daniels, 2012; Rush, Shelden, & Hanft, 2003). In the literature, coaching is defined as an interactive, dynamic, and ongoing process in which an expert models and supports a coachee to reflect on his/her practice (Gupta & Daniels, 2012; Neuman & Wright, 2010; Rush & Shelden, 2011). However, within the coaching concept, there is a great variety of models (Llyod & Modlin, 2012) and most approaches are not based on a specific theoretical framework (Isner et al., 2011).

Differentiations between in-service training approaches that offer scale-based, data-driven feedback and traditional approaches have not yet been accomplished. In the literature, it is assumed that scale-based formats strengthen the coaching process. Rating scales and data-driven feedback promote practitioner understanding, and the adoption and use of effective early childhood practice (Trivette, Raab, & Dunst, 2012a).

The review showed that in-service training is widely diverse with respect to training amount and duration. The aspect of training intensity has not been considered yet. A meta-analysis would be the next step to examine the influence of training formats and designs as well as training intensity.

Two additional potential confounding variables have appeared for in-service training effects on quality. First, several studies have evaluated in-service effects parallel to curriculum implementation. Klein and Gomby (2008) warn that the implementation of a new curriculum might be confounded with in-service training efforts. This makes it difficult to attribute the impact on quality ratings exclusively to the in-service training. In general, the impact of curriculum implementation on early childhood classroom quality is not clear. Barnett et al. (2008) indicate that a significant impact on quality ratings on the ECERS and CLASS took place through the implementation of the Tools of Mind curriculum (Bodrova & Leong, 2007). The Preschool Curriculum Evaluation Research study evaluated 14 curricula (PCER, 2008), but only the implementation of 2 language-specific curricula resulted in significant effects on quality ratings, but the implementation of the other curricula did not (e.g., "Let's begin with letters people" by Abrams & Company, 2000; "Literacy Express" by Lonigan, Clancy-Menchetti, Phillips, McDowell, & Farver, 2005; "Creative Curriculum" by Dodge, Colker, & Heroman, 2008 in North Carolina and Georgia). A meta-analytic examination of the moderator "curriculum implementation" would help to determine its influence on in-service training effects.

Second, the number of participants per approach varies. In particular, when large-scale studies or QRIS are launched, a great number of trainers and/or coaches are required. Heterogeneity in the coaching experience and the qualifications of trainers make it difficult to achieve a consistent service. Coaching interventions in QRIS or large-scale studies are more likely to vary than in smaller-scale studies (Tout, Isner, & Zaslow, 2011). Variations in professional development may mask the effectiveness of in-service programs. Further, the findings by Edgar (2008) demonstrate that the training and qualifications of coaches matter. Significant effects were gathered only for well-trained coaches and not through multipliers.

A meta-analysis would be the next step to investigate whether curricula confound professional development effects or whether large-scale professional development programs are less effective than small-scale programs that are piloted with a limited number of teachers.

2.3. Review of in-service training effects on language and literacy development

Providing an early foundation in school readiness skills has become a primary goal of many countries in order to decrease later school failure and dropout rates, in particular for children from families with a low socioeconomic background (National Evaluation of Sure Start (NESS) team, 2010). Recent research reinforces the importance of the use and understanding of vocabulary, complex oral language, phonological awareness, and letter knowledge on the development of later academic skills (National Institute for Literacy, 2008). However, there is a growing consensus that high-quality early child care can lay a strong foundation for disadvantaged children. Early childhood educators may not always have the formal qualifications or receive the educational preparation they need to provide highly effective classroom practice (Howes et al., 1992; Kontos, Howes, & Galinsky, 1996; Landry et al., 2009). Therefore, professional development may serve as a buffer for inadequate teacher preparation (Landry et al., 2009).

Several in-service training evaluations demonstrate the impact of effective instructional practice that promotes children's language and emergent literacy development in the early years (Beller & Beller, 2009; Buschmann et al., 2010b; Buysse, Castro, & Peisner-Feinberg, 2010; Wasik & Hindman 2011). A number of studies has examined the influence of variations in training delivery mode, including training formats and intensity (Flowers et al., 2007; Girolametto et al., 2003; 2004; 2007; Landry et al., 2009; 2011; Powell et al., 2010; Shidler, 2009). The impact of various in-service training initiatives on the language competencies of young children was investigated extensively. The findings of these programs on general language abilities are diverse. Positive effects were gained for scores assessing language comprehension skills (Beller, Merkens, & Preissing, 2007; Cordell, 2010; Onchwari & Keengwe, 2010), composite scores evaluating oral language proficiency (Barnett et al., 2008), and values describing language production and language processing skills (Schröder et al., 2011). No effects were found for scores estimating linguistic and pragmatic aspects (Beller & Beller, 2009), English language proficiency of Dual Language Learners (DLL; Buysse, Castro, & Peisner-Feinberg, 2010) or percentage of age-appropriate language skills (Wilcox-Herzog et al., 2013).

Through a systematic literature search, 55 studies with (quasi-)experimental, non-experimental or single-subject designs were found. The studies were published between 1990 and 2013 and conducted in the United States of America, Canada, Germany and Australia (see table 5). The review includes approximately one third unpublished studies ($n = 20$). The results provide an empirical foundation for the systematization of in-service training effects on the language and literacy development of young children.

Table 5. Review of studies measuring in-service training effects on language and literacy

Author	Year	Country	Pubtype	Design	Sample teacher	Sample children	WS	CS	ONL	COP	ON
Ahrens	2009	USA	unpub	CRT	6	24					x
Assel et al. (a)	2007	USA	pub	CRT	76	603	x				x
Assel et al. (b)	2007	USA	pub	CRT	76	603	x				x
Baker & Smith	1999	USA	pub	QNE	17	186				x	
Barnett et al.	2008	USA	pub	CRT	18	218	x			x	x
Beller & Beller	2009	GER	unpub	RCT	38	164	x				x
Beller et al.	2007	GER	unpub	QNE	31	155	x				x
Berufsbildungswerk	2011	GER	unpub	QNE	69	91		x			
Bos et al.	1999	USA	pub	QNE	31	324		x			x
Brown	2008	USA	unpub	QNE	3	46		x			x
Bryant et al.	2009	USA	unpub	RCT	108	236					x
Buschmann et al.	2010b	GER	pub	QNE	30	33		x			
Buyse et al.	2010	USA	pub	CRT	55	193	x			x	x
Clancy-Menchetti	2006	USA	unpub	CRT	40	500	x	x			x
Cohen et al.	2011	USA	pub	OGPP	6	75	x			x	x
Cordell	2010	USA	unpub	OGPP	3	34		x			x
Cusumano et al. (a)	2006	USA	pub	QNE	41	386		x			x
Cusumano et al. (b)	2006	USA	pub	QNE	41	386	x				
Domitrovich et al.	2007	USA	pub	QNE	20	275	x				x
Downer et al. (a)	2011b	USA	pub	CRT	242	1338	x		x		x
Downer et al. (b)	2011b	USA	pub	CRT	242	1338	x		x		
Flowers et al.	2007	CAN	pub	CRT	16	64		x			x
Gallagher et al.	2011	USA	pub	RCT	16	48					x
Girolametto et al.	2003	CAN	pub	CRT	16	64		x			x
Girolametto et al.	2004	CAN	pub	CRT	17	68		x			x
Girolametto et al.	2007	CAN	pub	RCT	16	64	x				
Graul & Zeece	1990	CAN	pub	OGPP	18	28		x			
Helmer et al.	2011	AUS	pub	OGPP	11	157	x		x		x
Hsieh	2005	USA	unpub	OGPP	5	58					x
Jackson et al. (a)	2006	USA	pub	QNE	55	230			x		
Jackson et al. (b)	2006	USA	pub	QNE	55	230			x		x
Jungmann et al.	2013	GER	pub	OGPP	27	263		x			
Kopacsi & Hochwald	1998	USA	unpub	QNE	21	1219	x				x
Krause et al.	2011	GER	unpub	QNE	118	418	x				x
Landry et al.	2009	USA	pub	CRT	262	1786			x		x
Landry et al.	2011	USA	pub	CRT	214	1571			x		
Lee	2010	USA	unpub	CRT	25	383		x	x		
Lonigan et al. (a)	2011	USA	pub	CRT	51	808	x				x
Lonigan et al. (b)	2011	USA	pub	CRT	51	808	x				
Mashburn et al.	2010	USA	pub	CRT	182	1165	x		x		x
McCutchen et al.	2002	USA	pub	QWM	44	492	x				x
McGill-Franzen et al.	1999	USA	pub	CRT	18	429	x	x			
Mohler et al.	2009	USA	pub	QNE	24	768	x	x			x
Morris et al.	2010	USA	unpub	QNE	52	623	x	x			x
O'Connor (a)	1999	USA	pub	QNE	10	154					x
O'Connor (b)	1999	USA	pub	QNE	17	311	x				
Onchwari & Keengwe	2010	USA	pub	QNE	44	626					x
Podhajski & Nathan	2005	USA	pub	QNE	86	126	x				x
Powell et al.	2010	USA	pub	RCT	88	759	x				x
Raspberry	2004	USA	unpub	CRT	4	38	x				
Rosof	2006	USA	unpub	OGPP	66	192	x			x	x
Rudd	2003	USA	unpub	CRT	59	121	x				x

(Table 5 continued)

Author	Year	Country	Pubtype	Design	Sample teacher	Sample children	WS	CS	ONL	COP	ON
Schröder et al.	2011	GER	unpub	OGPP	20	49	x				x
Shidler	2009	USA	pub	OGPP	24	360		x			x
Simon & Sachse	2013	GER	pub	QNE	49	146		x			
Sprague	2009	USA	unpub	CRT	160	840		x			
Tyler	2009	USA	unpub	QNE	4	170					x
Vernon-Feagans et al.	2010	USA	unpub	CRT	75	151			x		x
Wasik & Hindman	2011	USA	pub	CRT	30	541	x	x			x
Wasik et al.	2006	USA	pub	CRT	16	207		x			x
Wilcox-Herzog et al.	2013	USA	pub	OGPP	24	135				x	x

Note. USA=United States of America; CAN=Canada; GER=Germany; AUS=Australia; Pubtype=publication type; pub=published in journals or as book chapter; unpub=unpublished paper e.g., technical paper, dissertation or governmental report; design=study design; RCT=randomized control trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center to groups; QWM=quasi-experimental design with matching procedure/equivalent groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design; OGPP=one group pre-post measure; WS=workshop; CS=course and in-service with ongoing sessions; ONL=online training or blended learning; COP=community of learning or community of practice; ON=onsite support like coaching, consulting etc.

2.3.1. In-service training impacts on language development

Several studies evaluate the impact of different in-service programs on language skills, in particular *receptive vocabulary*, of young children. Some empirical evidence for beneficial effects on receptive vocabulary were found for traditional coursework (Graul & Zeece, 1990; McGill-Franzen et al., 1999) and for intensive professional development using onsite support (Barnett et al., 2008; Cohen, Kramer-Vida, & Frye, 2012; Jungmann, Koch, & Etzien, 2013; Landry et al., 2011; O'Connor, 1999; Shidler, 2009; Sprague et al., 2009; Wasik, Bond & Hindman, 2006; Wasik & Hindman, 2011). However, other investigations did not find a significant/satisfactory impact of teacher training on receptive vocabulary and complex language scores using different in-service training formats (Berufsbildungswerk Leipzig, 2011; Bryant et al., 2009; Buysse, Castro, & Peisner-Feinberg, 2010; Downer et al., 2011b; Gallagher, Abbott-Shim, & VandeWiele, 2011; Jackson et al., 2006; O'Connor, 1999; Powell, Diamond, Burchinal, & Koehler, 2010; Vernon-Feagans et al., 2010). However, the implementation of a specific curriculum (Barnett et al., 2008) and activities (O'Connor, 1999) within in-service training seem to be more promising than general instructional strategies (Buysse, Castro, & Peisner-Feinberg, 2010) or targeting reading instruction (Vernon-Feagans et al., 2010) to improve receptive vocabulary growth.

In addition to receptive vocabulary, a positive impact of diverse in-service approaches on the *expressive language* assessment scores of young children was confirmed. In particular, modern approaches such as online courses (Landry et al., 2009) and programs providing solely individual onsite support (Gallagher, Abbott-Shim, & VandeWiele, 2011; Hsieh, 2005;

Rudd, 2003) did foster expressive language abilities. Findings from intensive programs using a mixture of delivery formats were inconsistent (Assel et al., 2007; Barnett et al., 2008; Berufsbildungswerk Leipzig, 2011; Clancy-Menchetti, 2006; Cusumano et al., 2006, Jackson et al., 2006; Lonigan et al., 2011, Simon & Sachse, 2013, Weinstock et al., 2012). So far, beneficial effects have been gained for in-service training programs providing coursework and onsite support (Brown, 2008; Wasik, Bond & Hindman, 2006), the combination of online course and mentoring (Landry et al., 2011) as well as programs providing a workshop and mentoring (Assel et al., 2007, Lonigan et al., 2011). In contrast, traditional in-service programs using workshops as the primary delivery format seem to be ineffective in fostering expressive abilities in preschool and kindergarten students (Assel et al., 2007; Cusumano et al., 2006; Lonigan et al., 2011). Measuring the expressive vocabulary of infants and toddlers through a parent questionnaire, the study by Buschmann et al. (2010b) clearly indicates a significant growth through an interaction training course. In contrast, Krause and colleagues (2011) did not find a significant impact following an intense 115-hour team training approach on the expressive communication skills (teacher ratings) of kindergarten students with or without an immigration background.

Studies of linguistic approaches have examined the impact of in-service programs on the *utterances* of young children concerning verbal productivity, lexical diversity, syntactic complexity and responses at different abstraction levels. Fine-grained linguistic analyses of language development are based on video examinations of daily routines in center-based care. The majority of studies have investigated the effectiveness of different versions of the “Learning Language and Loving It” program from the Hanen Centre Toronto (Ahrens, 2009; Flowers et al., 2007; Girolametto et al., 2003; 2004; 2007). Further, Lee (2010) evaluated the impact of the MyTeachingPartner (MTP) program, Simon and Sachse (2013) the effectiveness of the Heidelberg Interaction Training (HIT) program, and Girolametto et al. (2012) the effect of the “ABC and Beyond” program on student utterances using video analysis. Findings from Simon and Sachse (2013) clearly indicate a significant growth in the *number of utterances* and *proportion of utterances* from children during book reading after the teacher interaction training at follow-up (10 months after the training), but not at posttest. Piasta and colleagues (2012) maintained their in-service training effects on the number of utterances and *length of utterances* after teachers enhanced the frequency of their communication-facilitating strategies through training. Investigating the “ABC and Beyond” literacy in-service program, using video analysis during book reading, Girolametto et al. (2012) found that children in the treatment group increased their *decontextual utterances* (i.e., extending story beyond text), and increased the *use of print referencing key words, alphabet letter names and sound awareness*. No effect was found for literal utterances (information from the text). Looking at *children’s responses* during book reading or play-dough activities, children from a “Learning

Language and Loving it” training group increased their number of responses to educators and peers (Girolametto et al., 2003; 2004; 2007). However, ratios and percentages of utterances, calculated as the proportion of all response opportunities, did not reach the significance level (Girolametto et al., 2004). In addition to the number and ratio of utterances, children’s *story comprehension utterances to educators* were coded. Response to story comprehension was examined, taking into consideration the level of abstraction during book reading. Abstraction levels distinguish between child response according to picture description (level 1), text and story awareness (level 2), background knowledge (level 3), and the integration (level 4) of analysis and evaluation of story events (cf. Flowers et al., 2007). Students from the training group increased their number of responses at level 3, a level with a more complex and higher abstraction level (Flowers et al., 2007; Girolametto et al., 2007), and decreased lower level responses (Flowers et al., 2007). Further, *lexical diversity* was investigated in terms of multiword utterances and different words. So far, there is limited evidence for impacts on multiword utterances (Girolametto et al., 2003) and for the number of different words (Lee, 2010; Piasta et al., 2012). However, the impact of different word utterances could not be confirmed by Girolametto and colleagues (2003). Up to now, only Lee (2010) and Simon and Sachse (2013) have investigated the impact of in-service programs on *syntax complexity*. Simon and Sachse (2013) examined the proportion of clauses in the utterances of children during book reading and found at least statistically considerable gains at follow-up. In the analysis by Lee (2010), syntax complexity of student responses was segmented into statement subordinations according to the clauses contained in the response. First, the findings indicated that most student responses did not include a complete sentence clause, but at least a single clause. Second, students in the MTP group gave a greater number of responses at all levels of syntactic complexity, but reached significance level only for low subordinate responses, an utterance that contains one clause (ibid., p. 123).

Comprehension skills, emphasizing the understanding of vocabulary, gestures, and complex language, are also evaluated as a target outcome of several in-service programs. However, the findings on the effectiveness of professional development on these skills are inconsistent (Assel et al., 2007; Schröder et al., 2011; Vernon-Feagans et al., 2010). Schröder and colleagues (2011) did not find a significant impact. The analysis by Vernon-Feagans et al. (2010) pointed out that intensive training is especially effective for children identified as struggling readers after controlling for covariates such as gender, mothers’ education, grade and race. Further, the study by Assel et al. (2007) investigated the impact of 2 different in-service training formats (with or without mentoring) used to implement 2 different curricula within 3 different settings on auditory comprehension skills in contrast to a control group. The findings showed that students from target classes had improved auditory comprehension abilities relative to controls. However, confounding effects were found for the mentoring con-

dition, site, and curriculum. Hence, it was not possible to attribute the impact of in-service training exclusively to mentoring, because the in-service training effect was moderated by type of program (ibid., p. 463).

The competence to repeat non-words is fundamental task for language acquisition. *Non-word repetition*, a short-term memory task, is known as one of the best predictors for learning new words and for the acquisition of vocabulary (Gathercole, 2006). Significant treatment effects on non-word repetition skills were found for interaction training programs (Jungmann, Koch, & Etzien, 2013; Simon & Sachse, 2013) and an extensive 2-year mentoring program to enhance the language and literacy environment (Clancy-Menchetti, 2006). In contrast, an intensive professional program did not lead to beneficial non-word repetition outcomes (Berufsbildungswerk Leipzig, 2011). Further, Lonigan et al. (2011) did not find significant treatment effects for the implementation of the Literacy Express Curriculum on non-word repetition, either for the mentoring or for the workshop condition. However, the authors expected the null results, because the Literacy Express Curriculum was not created to foster short-term phonological memory skills (ibid., p. 322).

2.3.2. In-service training impacts on literacy development

Emergent literacy, a precursor of formal reading, starts early in life and is highly associated with later school achievement (Strickland & Riley-Ayers, 2006; Whitehurst & Lonigan, 1998). Early literacy skills consist of alphabetic coding, including alphabetic knowledge, phonological and phonemic awareness, as well as print knowledge and concepts about print. The abilities a) to know the letters of the alphabet, b) to hear, identify and manipulate phonemes in words (phonemic awareness) and c) to distinguish sounds within words (phonological awareness) are known as strong school entry predictors of later reading skills (Adams, 1990; Honig, 1996; National Reading Panel, 2000; Riley, 1996; Strickland & Riley-Ayers, 2006). Intentional teaching is needed to increase the development of letter knowledge and phonological awareness (Strickland & Riley-Ayers 2006). Thus, there is a growing demand for capable teachers to provide high-quality early literacy experiences that might compensate for low levels of family literacy experiences.

Findings on composite scores on *emergent literacy* demonstrate that it is possible to improve these skills through a 2-day workshop (Podhajski & Nathan, 2005) or mentoring (Onchwari & Keengwe, 2010). Using a one-group-design, a descriptive analysis indicated that children made significant gains on phonological awareness, early literacy skills and vocabulary when

teachers learned how to implement the Early Reading Intervention through a 3-year teacher training program with workshops and coaching (Abbott et al., 2011).

Letter and alphabetic knowledge consists of a variety of sub-skills: the identification and naming of letters, spelling of words and the knowledge how to write letters (cf., Wood & McLemore, 2001). Children named more letters when their teachers received mentoring or individual onsite support (O'Connor, 1999; Gallagher, Abbott-Shim, & VandeWiele, 2011) or participated in intensive programs (McGill et al., 1999, Sprague et al., 2009), online courses (Landry et al., 2011), college courses (Jackson et al., 2006), or even in a single workshop (O'Connor, 1999). In contrast, no significant changes occurred for other intensive formats on letter naming or alphabet knowledge (Jackson et al., 2006; Mohler et al., 2009; Wasik, Bond, & Hindman, 2006; Wasik & Hindman, 2011), not even when classified into first and second language acquisition (Buysse, Castro, & Peisner-Feinberg, 2010). Further, the results on letter naming fluency (Tyler, 2009) and letter sounds (Landry, 2011) were unrewarding. Differentiation between upper and lower case letters revealed significant gains and group contrasts with lower case letters, but no effects for the interaction of group and time was found (Gallagher, Abbott-Shim, & VandeWiele, 2011). The ability to name letters and read words aloud from a list, measured with the Woodcock Johnson letter-word identification test, was tested as a distal in-service training outcome in some studies. O'Connor (1999) as well as Vernon-Feagans and colleagues (2010) found positive effects for typical learners and at risk/struggling children on letter-word identification skills after teachers participated in professional development. Specific-curriculum was used by Vernon-Feagans and colleagues (2010) with target reading instruction with a one-on-one intervention and by O'Connor (1999) with specific activities in comparison to more general approaches that linked literacy to early childhood classrooms. No effects were found by Powell et al. (2010) and by Barnett et al. (2008).

Phonological awareness is generated in early childhood and consists of several components (e.g., rhyming, segmentation, blending, and phoneme awareness) that are sorted into a continuum of complexity (Chard & Dickson, 1999). Several studies provide significant proof that professional development can foster less complex *phonemic awareness* abilities (Buysse, Castro, & Peisner-Feinberg, 2010; Gallagher et al., 2011; McGill-Franzen et al., 1999; Wasik & Hindman, 2011). However, Landry et al. (2011) did not find evidence for training effects on complex phonological awareness assessments (Landry et al., 2011).

The ability to produce and identify *rhyming* words is a prerequisite for the development of more complex phonological awareness (Torgesen & Mathes, 1998). In particular, the ability to rhyme is at the less complex end of the continuum (Chard & Dickson, 1999) and can easily be fostered through rhyming songs. Several studies demonstrate the impact of different in-service professional development programs on rhyming. Children profited significantly when

their teacher participated in a 4-month program with coursework and onsite support (Brown, 2008; Cusumano et al., 2006). Further, distal training effects on rhyming abilities occurred when teachers received 2.5 months of individual onsite support (Hsieh, 2005). Mohler et al. (2009) showed that children improved their rhyming scores, rated on the Emergent Literacy Survey, when teachers participated in a 2-year intensive in-service program called the Pre-school Literacy Project. Investigating the impact of 2 different intensity levels of in-service programs, O'Connor (1999) found significant effects on rhyming skills at least for at risk children in the workshop condition, but not for the intensive program. The impact of the Nuestros Ninos in-service program on Latino Dual Language Learners was tested in the first and second language (Buysse, Castro, & Peisner-Feinberg, 2010). Significant effects were reported for rhyming abilities in Spanish, but not for the second language, English (ibid., p. 201).

The ability to isolate and pronounce the (initial) sound of words is a key competency at the end of kindergarten. There is some evidence that supports the assumption that the capacity to pronounce the *beginning sound* of a picture is fostered indirectly if teachers receive mentoring (Gallagher, Abbott-Shim, & VandeWiele, 2011) and intensive training (Mohler et al., 2009). Further, children had higher values in a sound identification task (where they were asked to produce the sounds of alphabetic letters) after teachers participated in an intensive in-service program (Bos et al., 1999). Children had higher initial sound fluency when their teacher participated in a 4-month training program that demonstrated how to embed instruction and elicit responses that foster literacy development in daily routines (Brown, 2008). However, no in-service training effects occurred for a similar initial sound matching task asking children to point to a picture that begins with the same sound as the target picture (Powell et al., 2010).

Word awareness skills, a critical prerequisite of early reading achievement, were examined by a test for *blending*, where children were asked to combine word sounds to make a new word. Several professional development programs were made to enhance the blending abilities of young children. The findings were inconsistent when all professional development programs implemented specific teaching practice through curricula. O'Conner (1999) reported positive effects on blending skills for typical learners and at risk children by providing individual support or a single workshop for teachers to implement the Ladders to Literacy Program. Mohler and colleagues (2009) and Powell et al. (2010) provided some evidence for the promotion of blending abilities through intensive in-service training. However, direct associations were not found for children after their teachers participated in single workshops (Lonigan et al., 2011) or in intensive in-service programs using different delivery formats (Clancy-Menchetti, 2006; Downer et al., 2011b; Lonigan et al., 2011; Rosof, 2006).

Phonemic segmentation, the ability to break down words into individual sounds, was a favored outcome of a limited number of studies. Significant treatment effects occurred for tradi-

tional in-service training and intensive professional development formats (O'Connor, 1999) as well as the intensive Preschool Literacy Project providing an online course and coaching to implement enriched literacy practice in classrooms (Mohler et al., 2009). Findings from Baker and Smith (1999) suggest a positive impact on segmentation benchmarks. After providing an intensive in-service program for teachers to enhance literacy teaching for Hispanic DLLs, treatment effects were only found in the onset segmentation and matching task in English, but not in Spanish (Buysse, Castro, & Peisner-Feinberg, 2010).

Only a handful of studies have evaluated the impact of in-service programs on phonemic awareness skills, measured with the *elision* task from the PRE CTOPPP (Clancy-Menchetti, 2006; Rosof, 2006; Downer et al., 2011b; Lonigan et al., 2011). In the elision task, children determine how the meaning of a word changes when one of its component sounds is removed. With the exception of the exploratory study by Rosof (2006), the descriptive data of the other evaluations (Clancy-Menchetti, 2006; Downer et al., 2011b; Lonigan et al., 2011) indicate a growth in elision skills. These results need to be discussed with caution, because only descriptive statistics were available.

Print knowledge and print awareness are key elements of early literacy development and strong determinates of early reading achievement (Adams, 1990; Justice & Ezell, 2001). Children understand what print looks like, are aware that print has meaning, and master the understanding of the concepts of words and letters prior to learning to read (Justice & Ezell, 2001; Strickland & Schickedanz, 2004). Research demonstrates that the print knowledge skills of young children can easily be improved through professional development. Learning how to use environmental symbols, how to make reading and writing meaningful and how to teach print that carries over into beneficial child outcomes was successfully delivered through traditional in-service training (McGill-Franzen et al., 1999; Rasberry, 2004), intensive in-service programs (Abbott et al., 2011; Downer et al., 2011b; Lonigan et al., 2011; Mohler et al., 2009; Powell et al., 2010), and online formats (Landry et al., 2009; 2011). However, not all in-service programs were able to considerably change print awareness (Clancy-Menchetti, 2006; Lonigan et al., 2011).

Only a handful of studies have evaluated the impact of professional development on *(pre-)reading skills*. Teaching how to read was implemented mainly through workshops and individual onsite support (Helmer et al. 2011; McCutchen et al., 2002; Kopacs & Hochwald, 1998). McCutchen and colleagues (2002) did not find significant differences in word reading abilities between experimental and control groups at the end of the kindergarten year. However, the contrast appeared significant when the actual use of phonological teaching strategies was considered. Descriptive results suggest that there were gains for kindergarten children and first grade students at classroom level in group reading when teachers implemented a software program for reading support (Helmer et al., 2011). After teachers implemented

a multicultural curriculum in a kindergarten, descriptive findings indicated that the children in the experimental group outperformed the children in the control group on reading assessment (Kopacsi & Hochwald, 1998). Further, significant effects were found for children in the Individual Learning Intervention group, a mentoring program to enhance the quality of the learning environment (Gallagher, Abbott-Shim, & VandeWiele, 2011). Vernon-Feagans et al. (2010) demonstrated that struggling readers profited most when teachers learned how to implement Target Reading Instruction through an online course and individual support.

Throughout, positive effects have been found for in-service programs on *dictation* scores in a limited number of investigations. Students whose teachers received professional development through traditional workshops (O'Connor, 1999), coursework (Bos et al., 1999; Jackson et al., 2006), intensive individual training (O'Connor, 1999) or mentoring (Gallagher, Abbott-Shim, VandeWiele, 2011) outperformed students in the control condition. However, students whose teachers received additional mentoring in the HeadsUp! Reading (HUR) course made non-significant gains in comparison to control children, while the other students whose teachers only participated in coursework made significant gains (Jackson et al., 2006).

A limited amount of studies have evaluated the impact of in-service training on early and *emergent writing* skills. In particular, literacy-focused in-service programs targeting literacy instruction and the print environment have shown positive effects (McGill-Franzen et al., 1999; Mohler et al., 2009; Powell et al., 2010). However, non-specific professional development programs targeting several aspects of language and literacy did not improve the emergent writing skills of young children significantly (Krause et al., 2011).

2.3.3. Differential effects of in-service training for Dual Language Learners and children at risk

Some studies have analyzed the impact of in-service training for Dual Language Learners (DLL) or English Language Learners (ELL) and children at risk. This section brings together the findings for the different subgroups. The retrospective analysis by Fitzer (2009) using Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K) data from 3,762 teachers and 21,260 Hispanic kindergarten students demonstrated that teachers' professional development had an impact on the literacy achievement of kindergarten Hispanic ELLs. The findings clearly indicated that teachers who attended 3 or more in-service training sessions and teachers who completed a college course were able to affect the literacy scores of Hispanic ELLs significantly. Differential analysis of the impact of in-service programs on the second language abilities of *Dual Language Learners* have been carried out especially by German researchers, using quasi-experimental and non-experimental designs (Beller & Beller, 2009;

Beller, Merken, & Preissing, 2007; Jungmann et al., 2013; Krause, Kofler, & Hofbauer, 2011; Krause & Kofler, 2011; Schröder et al., 2011; 2012). In particular, positive impacts on a variety of *language and literacy outcomes in the second language* for DLLs were found for in-service programs that qualify teachers to embed specific language-promotion and interaction-fostering strategies in daily routines (Beller & Beller, 2009; Beller, Merken, & Preissing, 2007; Jungmann et al., 2013; Schröder et al., 2011; 2012). After implementing an elaborated linguistic style through intensive training, DLLs in the training condition made greater gains in language production and processing (Schröder et al., 2011; 2012). However, the improved growth rate could not be maintained at follow-up, 6 months after training. A comprehensive program, combining (a) interaction training, (b) some sessions on phonological awareness training and (c) a program for embedded strategies for dual language learners, was evaluated by Jungmann, Koch, and Etzien (2013). Their findings demonstrated a significant reduction rate - from 52 % to 48 % - in DLLs identified as having negligible and very limited language proficiency skills. Further, differential analysis of DLLs identified as typical learners indicated the same gain rates as monolingual children, but very low-skilled DLLs had lower growth rates in language comprehension, processing and production (ibid., p. 118). Enhancing the use of language-supporting strategies and linguistic styles in daily routines was focused on in the intensive program by Beller and Beller (2009). Findings showed that Turkish and German preschoolers benefited equally from teacher training and improved teaching practice. Differential analysis of Turkish and German preschoolers, identified as low achievers at pretest, showed that significant improvements appeared only for the group of Turkish students. A similar professional program for teachers in infant and toddler classrooms was conducted by Beller and colleagues (2007). Comparing the language ability scores of Turkish, Arabic and German infants and toddlers, children from all ethnic groups profited equally from a teacher intervention that implemented language-promoting strategies in daily routines (Beller, Merken, & Preissing, 2007). Taking a closer look at the effective programs that support second language acquisition, it becomes apparent that these interaction programs implement language and interaction-promoting strategies using videos or transcript analysis for reflection, role-play, and performance-based assignments. In contrast to interaction training, the knowledge-based program “Sprachberater”, which includes 115 hours of team coaching and workshops, did not show significant changes in language outcomes for DLLs (Krause, Kofler, & Hofbauer, 2011; Krause & Kofler, 2011).

In addition to second-language skills in English, only 3 studies have examined the effects of professional development on the *first language abilities in Spanish* of dual language learners in the United States (Barnett et al., 2008; Buysse, Castro, & Peisner-Feinberg, 2010; Landry et al., 2011). Differentiating between the English and Spanish abilities of Latino children in center-based care, Buysse and colleagues (2010) reported significant gains on some phono-

logical awareness task scores, measured in Spanish and English, for children whose teachers had participated in the Nuestros Niños program. However, the effects could not be seen in other test scores. Oral language proficiency in Spanish was investigated in a subgroup of children while their teachers received training to implement the Tools of Mind curriculum (Barnett et al., 2008). The results showed that the children in the Tools of Mind condition had higher oral language proficiency in Spanish. Further, their receptive language scores (PPVT) in English were predicted by the primary home language and in particular children whose home language was English had higher scores (ibid.). Comparing 2 different lengths of in-service training, Landry et al. (2011) reported that program effects were dependent upon the language of testing. English Language Learners with lower vocabulary scores benefited more when their teachers participated in a second year of professional development (p. 992). The current state of research, especially the small number of studies, is part of the general problem in measuring the first language abilities of children. So far, only an unrewarding number of instruments are available to measure first language abilities.

Differential analysis for children diagnosed as *low skilled or at risk of developmental failure* has been carried out in a handful of studies (Abbott, Atwater, Lee, & Edwards, 2011; Beller & Beller, 2009; Jungmann, Koch, & Etzien, 2013; O'Connor, 1999; Simon & Sachse, 2013; Vernon-Feagans et al., 2010). Comparing a traditional in-service program and an intensive training format, O'Connor (1999) found significant treatment effects for both programs on several language outcomes for children with lower skill levels and students with disabilities. However, descriptive analysis indicated that at risk children in the intensive professional development condition could not catch up with typical learners in blending, segmenting, word-letter-identification and dictation (ibid., p. 208). One-on-one target reading instruction was implemented for focal children through a distance technology and in-service coaching model (Vernon-Feagans et al., 2010). Controlling for pretest differences, struggling readers in the training condition improved their word attack, letter-word identification, passage comprehension, and spelling sound scores significantly more than struggling readers in the control condition. This result could not be confirmed for expressive vocabulary. Further, there is some evidence that struggling readers catch up with typical learners from the in-service training condition in letter-word identification, word attack, and spelling of sound scores (ibid., p. A-5). Children with low levels of semantic and vocabulary skills (lowest third) were investigated separately by Simon and Sachse (2013). Their findings demonstrated short-term effects in video ratings of lower skilled students on language production, including raw scores of constituents, prepositions and subjects, immediately after interaction training for teachers. Further, a long-term impact was found for the production of subjects and prepositions 6 months later. However, the group of low achievers in both the training and control conditions had similar growth rates on standardized tests of vocabulary, grammar, comprehension and lan-

guage processing. The findings by Jungmann and colleagues (2013) showed a significant decrease in the percentage of children who did not reach the language assessment benchmark in the in-service training condition. Low achievers from the training group made larger gains in language comprehension, language and phonological processing than low achievers in the control condition. While Beller and Beller (2009) did not find a significant difference in general language abilities between intervention and control groups after implementing language-promoting strategies in classrooms, the differential analysis on mono- and bilingual children identified as at risk demonstrated a statistically considerable reduction in scores significantly below the age norm, at least for bilingual Turkish students in the in-service training condition. A 3-year professional development program with 104 training hours of workshops and coaching was investigated by Abbott and colleagues (2011). After 3 years of training to implement the Early Reading First Intervention, 100 % of children reached the benchmark (standard score of 90) on the TOEPL Early Reading Index, while only 54 % did in the second year.

After splitting the study participants into specific subgroups (e.g., mono- and bilingual children, typical and struggling learners), heterogeneous results on the effectiveness of in-service programs and implemented interventions/strategies have been found. This kind of analysis provides more specific information on the response to interventions of a subset of children within studies. So far, the small but growing body of evidence suggests that responses to interventions/teacher training vary significantly. Splitting the participants into typical and struggling learners, low, middle and high skill levels, as well as using age-appropriate benchmarks/cut-offs, would make it possible to investigate the compensatory effects of professional development. Compensatory effects are reached when struggling/low skilled students are able to catch up with the reference groups of typical learners or at least middle-skilled students. Subgroup analyses within systematic reviews or meta-analyses are usually uncommon, because of insufficient information in primary investigations (Higgins & Green, 2011). Therefore, more studies analyzing subsets of participants are needed to provide detailed information on responses to intervention/treatments in order to better understand compensatory effects.

2.3.4. Evidence base on in-service training effects on language and literacy

To sum up, there is a large evidence base investigating the impact of in-service professional development programs on language and literacy outcomes for children in center-based care. A few studies were carried out between 1998 and 2013, while most of the studies were published in the last 5 years. More than 60 different professional development approaches have

been evaluated on their effectiveness to foster language and literacy outcomes. It has become apparent that there is a tendency for in-service training to have larger effects on some outcome constructs than on others. Further, the magnitude of the evidence base varies tremendously (see table 6).

Table 6. Evidence base of in-service training effects on language and literacy development

Outcome	Magnitude of evidence base	Trend	Tendency of in-service effect
Receptive language	Large	0	Inconsistent, approximately equal numbers of studies demonstrates in-service training effects or no results
Expressive language	Large	-	Somewhat more studies indicate no results
Utterance	Medium	+	Throughout positive in-service training effects
Comprehension	Small	-	Small number of studies indicates no results
Non-word repetition	Medium	0	Inconsistent; approximately equal amount of studies demonstrates in-service training effects or no results
Alphabet knowledge	Large	0	Inconsistent, approximately equal amount of studies demonstrates in-service training effects or no results
Phonological awareness	Small	-	Small amount of studies indicates no results
Phonemic awareness	Small	+	Small amount of studies indicates in-service training effects
Rhyming	Medium	+	Some studies demonstrate in-service training effects
Beginning sound	Small	+	Small amount of studies indicates in-service training effects
Blending	Medium	0	Inconsistent; approximately equal amount of studies demonstrates in-service effects or no results
Phonemic segmentation	Small	+	Small amount of studies indicates in-service training effects
Elision	Small	+	Small amount of studies indicates in-service training effects
Print knowledge	Large	+	Large evidence base demonstrates in-service training effects
Early reading	Small	+	Small amount of studies indicates in-service training effects
Dictation	Small	+	Small amount of studies indicates in-service training effects
Writing	Small	+	Small amount of studies indicates in-service training effects

Note. Large=> 10 treatments; medium=> 6 treatments; small=> 6 treatments; + =positive trend; 0=no trend; -=negative trend.

So far, there is a large and solid evidence base with more than 10 investigations on the impact of in-service training on receptive and expressive vocabulary, letter naming and alphabet knowledge as well as for print knowledge. Some investigations (6 to 10 studies) were conducted to examine the effectiveness of professional development programs on non-word

repetition abilities, rhyming and blending skills as well as different aspects of utterances. A limited number of studies (< 6) analyzed in-service training effects on language comprehension, total scores for phonological or phonemic awareness, beginning sound identification, phonemic segmentation, elision, and early reading abilities. In particular, findings on those outcomes with a small evidence base need to be interpreted with caution and more research is needed. As for trends in the results of research on a single aspect of language or literacy scores, a large evidence base demonstrates that print knowledge can be fostered through professional development. Summarizing findings from studies examining in-service training effects on expressive vocabulary scores, somewhat more studies suggest that there is no effect than those that do. An inconsistent evidence base also appeared for receptive vocabulary development and letter identification and alphabetic knowledge abilities, with an equal number of studies showing in-service training effects or no results.

Some studies (n = 6) show the impact of in-service programs on rhyming skills and a broad variety of child utterances and responses, measured through video-analysis. There are inconsistent findings for non-word repetition and blending abilities, with approximately equal numbers of studies indicating some in-service training effects or no changes.

The findings on professional development effects on language comprehension, phonological and phonemic awareness composite values, beginning sound identification, phonemic segmentation, elision, early reading abilities and dictation abilities should be conservatively interpreted due to a very limited evidence base. Thus far, the results from 5 or fewer studies have suggested a positive impact of in-service programs. In contrast, no in-service training effects have been found for language comprehension and phonological awareness composite scores yet. Potential explanations for the inconsistent results and trends are multifaceted.

First, utterances and responses are very fine-grained units of language and might be more aligned to treatments. Fukkink and Lont (2007) found in their meta-analysis that studies which aligned their measures more closely to treatments obtained larger effects (see Fukkink & Lont, 2007). This might explain why positive in-service training effects were found throughout for measures of utterances and responses conducted through video analysis.

Second, inconsistencies in the findings appear to be due to the magnitude of genetic and environmental influences on different outcome constructs. Samuelsson et al. (2007) found substantial genetic influences on phonological awareness and verbal memory including non-word repetition. Strong environmental influences appeared for vocabulary and print knowledge. Research has shown that the environment plays a central role in language learning in children with higher non-word repetition abilities, whereas the language and reading development of children with poor non-word repetition skills at an early age is less affected by the environment (Bishop, 2001; Bishop, Adams, & Frazier Norbury, 2004). The ability to

repeat non-words is linked to speech and language acquisition, phonological awareness and early literacy acquisition (Hartmann, Rvachew, & Grawburg, 2008; Hoff, Core, & Bridges, 2008). In particular, it is highly correlated to vocabulary growth, at least for typically developing children (Gathercole, 2006), and very important for the successful completion of phonemic blending tasks (Brady, 1997; Hartmann, Rvachew, & Grawburg, 2008). In particular, low levels of non-word repetition skills as a covariate for receptive and expressive vocabulary growth and the development of several aspects of phonological awareness should be examined in further investigations.

Third, the fact that, regardless of any genetic influence, poor and high levels in language and reading abilities are also caused by factors in the child's environment is uncontroversial (Bishop, Adams, & Frazier Norbury, 2004; Olson et al., 2007). In particular, reading is a skill learned by instruction (Bishop, Adams, & Frazier Norbury, 2004) and Samuelsson and colleagues (2007) suggest that sufficient supplemental instruction is needed to move students' competencies in pre-reading and emergent reading to higher levels. Instructional practice, activities and curricula are implemented in early childhood classrooms through professional development. Landry et al. (2006) demonstrate that the impact of in-service programs is moderated by curricula. In particular, the impact is larger when a research-based early literacy curriculum is implemented in comparison to philosophy-based approaches (e.g., High/Scope or Creative Curriculum) or to generic programs (ibid., p. 317).

Fourth, discussing the inconsistent findings on alphabetic knowledge, it can be hypothesized that the content of in-service programs influences their training impact. The meta-analysis by Piasta and Wagner (2010) shows small-to-moderate instructional effects on alphabetic knowledge skills in preschool or kindergarten students. However, the impact differed according to the instruction provided, while small group interventions yielded larger effects than larger groups (ibid., p. 8). Inconsistencies in the impact of in-service training might occur because of different instructional practices that are implemented through teacher training. The type of instruction or curriculum content in professional development raises another critical issue and are potential moderating factors that could influence in-service training effectiveness. The content of in-service training matters according to Buysse, Winton, and Rous (2009), but more research is needed that evaluates the effectiveness of single features of in-service programs even though there is already a large number of studies investigating language and literacy development.

Fifth, a limited amount of studies has carried out subgroup analyses for struggling and typical learners. The findings indicate that children with different starting levels respond differently to interventions. Starting competency level might operate as a moderator for the effectiveness of in-service programs and needs to be considered in further research. In addition to different

competence levels, subgroups of mono- and bilingual children might explain inconsistent findings.

In summary, several questions on potential effect modifiers appeared during the review process. More meta-analytic examinations would provide further insights why some studies demonstrate effects and some do not. Features of professional development, the implementation of specific curricula, strategies or activities brought about through training, or focusing on a specific group of children might all operate as moderators for in-service training effects and need to be further investigated.

2.3.5. Challenges measuring in-service training effects on language and literacy

There is a substantial research base evaluating initiatives to strengthen early childhood teachers' intervention practice to foster children's language and literacy abilities at an early age. However, it is important to notice that most of the studies also pertain to the implementation of specific tools/strategies or curricula in classrooms (Zaslow et al., 2010b). Klein and Gomby (2008) remark that only a limited amount of studies are designed to hold curricula constant to make it possible to examine professional development effects without confounding them with *curriculum impact*. The findings by Assel et al. (2007) implementing 2 different curricula ("Let's Begin With Letters People" and "Doors to Discovery") through mentoring confirm this assumption. The impact of mentoring was less clear and seemed to depend on curriculum and type of language or literacy skill.

Further, language and literacy development encompasses multiple skills, including oral language productivity, vocabulary, language comprehension, phonemic and phonological awareness, alphabet and print knowledge and early reading and writing skills (Chard & Dickson, 1999; Jude, 2008; Grimm & Weinert, 2008; Weinert, Doil, & Frevert, 2008). Since language and literacy are key competencies and reliable predictors of later school success (cf. Ebert & Weinert, 2013), there is a growing demand for qualified teachers that are able to execute a variety of strategies to support the full range of language and literacy competencies. Such strategies include, for example, shared book reading, purposeful questioning, arranging book and literacy areas, increasing alphabetic and print teaching and materials and assessing and monitoring individual growth. The review demonstrates that professional development programs usually emphasize more than one of the above mentioned strategies (Dickinson & Brady, 2006; Zaslow et al., 2010a; 2010b). To capture the multiple facets of language and literacy development, it is appropriate to implement *multiple strategies within*

training, but this makes it difficult for research on training effectiveness to disentangle which strategy is related to which outcome (Zaslow et al., 2010b).

The degree to which teachers implement new strategies or curricula in classrooms also seems to matter. *Fidelity* was measured in a small amount of studies (cf. Barnett et al., 2008; Cusumano et al., 2006; Domitrovich et al., 2007; Jungmann et al., 2013; Vernon-Feagans et al., 2010; Wasik & Hindman, 2011). However, only 3 studies have investigated the confounding role of fidelity and found some small associations between language scores and fidelity to curriculum or specific language-promoting strategies (Cusumano et al., 2006; Jungmann et al., 2013; Wasik & Hindman, 2011).

In addition to several techniques and strategies that are delivered through in-service interventions, a great variety of *delivery formats and methods* are used to support teachers implementing new strategies (e.g., traditional workshops or courses, online formats, intensive programs with onsite support). So far, a handful of studies has compared the effectiveness of different delivery modes and indicate that this makes a difference to children's language and literacy achievement (Cusumano et al., 2006; Downer et al., 2011b; Lonigan et al., 2011; O'Connor, 1999; Powell et al., 2010). In the study by Shidler (2009), teachers received different amounts of coaching over a period of 3 years, ranging from 60 to 469 hours. The findings suggest that *training intensity* might influence the effectiveness of coaching.

In summary, several questions on potential effect modifiers appeared during the review process. It became apparent that the evidence base differs according to outcome construct. For meta-analytic examination, it is important to distinguish in-service training effects on different outcome domains. Features of professional development, training intensity, the implementation of specific curricula, strategies or activities brought about through training, or focusing on a specific group of children might all operate as moderators for in-service training effects and need to be investigated through meta-analytic techniques.

2.4. Review of in-service training effects on social skills and reduced problem behavior

Positive relationships and social interactions with peers are associated with language acquisition, advanced cognitive development, and the development of more sophisticated social behavior (Beelmann, 2004; Hendrickson, Gardner, Kaiser, & Riley, 1993; McEvoy, Odom, & McConnell, 1992). Children with high levels of problem behavior are at risk of peer rejection and later antisocial behavior in adolescence (Campbell et al., 2006; Tremblay, 2000; Tremblay et al. 2004; Webster-Stratton, 1997). Children's use of prosocial behavior may enhance their social status in the group and serves as a protective factor against ongoing rejection by peers (Coie & Kupersmidt, 1983; Crick, Casas, & Mosher, 1997; Dodge, 1983; Ladd, Price, & Hart, 1988). Nevertheless, a number of young children possess conduct difficulties or exhibit problem behavior (cf. Berg & Tisdale, 2004; Ihle & Esser, 2002; Lavigne et al., 1996; Webster-Stratton, 1997). Children who enter school with poor social skills or behavioral difficulties are at risk of inadequate behavioral adjustment and less classroom participation that leads to poor achievement (Raver & Knitze, 2002). This makes it imperative for teachers to implement effective behavioral management techniques (Filcheck, McNeil, Greco, & Bernard, 2004). Whole-class interventions might be more time and cost-effective than child-centered approaches using individual behavior plans. This presumption implies the need for an (ecological) approach that works directly with preschool teachers to improve their skills in managing behavior in classrooms to prevent problems and promote prosocial skills while considering the role of the environment and persons in the environment in shaping child behavior (Zirpoli & Melloy, 1993).

A systematic literature search in the electronic data bases and renowned journals was conducted to find studies that have investigated the impact of professional development on social-emotional development and problem behavior of children in center-based care. Twenty-eight studies with 29 treatments, conducted between 1992 and 2012, evaluated the impact of in-service professional development programs for preschool teachers on improved social skills or reduced problem behavior in young children (see table 7). Approximately 2 thirds of the reports ($n = 19$) were published in peer review journals or books and one third ($n = 9$) of the studies were grey literature (e.g., research reports, dissertations, reports or conference papers). Most of the included reports were quasi-experimental studies conducted in the United States. Small and large-scale studies were included in the systematic review. The total samples of teachers in the investigations varied from 3 to 750 participants at the beginning of the studies. The most common professional development delivery approaches to foster social skills and inhibit negative behavior used (a) a combination of workshops followed by on-site support, (b) a mixture of foundation workshop, coursework and individual onsite support,

(c) solely onsite support, and (d) courses. Several other training formats were offered such as workshop only (Wood, 1997), access to online resources exclusively and with onsite support (Driscoll et al., 2011) as well as intensive formats combining courses and technical assistance (Weinstock et al., 2012) or workshop, community of practice meetings and individual support (Barnett et al., 2008).

The *duration* of the in-service programs for preschool or kindergarten teachers varied considerable. Seven programs were short-term approaches with less than 2 months of training and 8 programs consisted of in-service training approaches with at least 9 training months.

The *content* of the in-service programs focused on social-emotional and behavioral aspects of young children, the implementation of general or specific curricula, quality improvement or more general topics concerning children's needs and learning processes, enriched learning environment and developmentally appropriate practice. Several studies evaluated the impact of preschool teachers' participation in in-service professional development programs on pro-social and negative behavior in preschool-aged children as a result of changed teacher behavior. Some in-service training approaches directly fostered social, emotional and behavioral skills through (a) the implementation of curricula, such as Tools of the Mind (Tools), Promoting Alternative Thinking Strategies (PATH) or I Can Problem Solve (ICPS) (Barnett et al., 2008; Domitrovich, Cortes, & Greenberg, 2007; Vestal & Jones, 2004), (b) the introduction of positive discipline techniques and behavior support (Sperry, 1999; Thanasetkorn, 2009), (c) the establishment of emotional and behavioral management in classrooms (Baker-Henningham, Scott, Jones, & Walker, 2012; Demchak, Kontos, & Neisworth, 1992; Shernoff & Kratochwill, 2007; Snyder et al., 2011) or (d) the accomplishment of multi-level models that included functional assessment (Stoiber & Gettinger, 2011), a curriculum (Baker-Henningham, Walker, Powell, & Meeks Gardner, 2009) or a parental component (Filcheck, McNeil, Greco, & Bernard, 2004) in addition to behavior management. Nevertheless, other professional development programs existed to promote social abilities and to diminish problem behavior in young children in center-based care. For example, quality improvement initiatives (Bryant et al., 2009) and in-service approaches that facilitated developmentally appropriate practice (cf. Gallagher, Abbott-Shim & Vande-Wiele, 2011; Rhodes & Hennessy, 2000) were also implemented. Further, training programs focused on language and literacy strategies were established to affect social-emotional and behavioral skills indirectly through the introduction of child-centered social interaction and language-promoting strategies in classrooms (Cordell, 2010; Girard, Girolametto, Weitzman, & Greenberg, 2011; Landry et al., 2009).

Table 7. Review of studies measuring in-service effects on social skill and problem behavior

Author	Year	Country	Pubtype	Sample	Design	Topic	Content	WS	Course	Online	COP	Onsite	Duration
Baker-Henningham et al.	2012	Jamaica	pub	73	CRT	social	BM	x				x	5.00
Baker-Henningham et al.	2009	Jamaica	pub	27	CRT	social	BM+C+M	x				x	7.00
Barnett et al.	2008	USA	pub	18	CRT	curriculum	C	x			x	x	9.00
Bryant et al.	2009	USA	unpub	108	CRT	quality	Q					x	9.00
Cordell	2010	USA	unpub	3	OGPP	curriculum	C		x			x	3.00
Demchak et al.	1992	USA	pub	9	MBL	social	BM					x	0.20
Domitrovich et al.	2007	USA	pub	20	CRT	social	C	x				x	9.00
Driscoll et al.	2011	USA	pub	252	QNE	social	S			x		x	n.e.i.
Driscoll et al.	2011	USA	pub	252	QNE	social	S			x			n.e.i.
Everston & Smithey	2000	USA	pub	46	QNE	social	BM	x				x	5.00
Filcheck et al.	2004	USA	pub	3	MBL	social	BM+M	x				x	1.00
Gallagher et al.	2011	USA	pub	16	RCT	general	G					x	9.00
Girard et al.	2011	Canada	pub	17	CRT	social	S	x				x	1.50
Graul & Zeece	1990	USA	pub	18	QNE	general	P		x				1.75
Hemmeter et al.	2011	USA	pub	4	MBL	social	S	x		x			4.00
Jo Rodriguez et al.	2009	USA	unpub	3	MBL	social	BM					x	3.00
Landry et al.	2009	USA	pub	750	CRT	language	G	x	x			x	12.00
Lloyd & Millenky	2011	USA	unpub	91	CRT	social	BM+SM+CCC	x	x			x	9.00
Morris et al.	2010	USA	unpub	51	CRT	curriculum	BM+SM+CCC	x	x			x	9.00
Rhodes & Hennessy	2000	UK	pub	33	QNE	general	G		x				6.00
Sherhoff & Kratochwill	2007	USA	pub	8	CRT	social	BM					x	1.25
Snyder et al.	2011	USA	pub	28	QNE	social	BM		x			x	3.00
Sperry	1999	USA	unpub	6	MBL	social	S	x				x	1.25
Stoiber & Gettinger	2011	USA	pub	70	CRT	social	BM+FA	x				x	4.50
Thanasetkorn	2009	Thailand	unpub	20	QNE	social	S	x				x	1.50
Vestal & Jones	2004	USA	pub	11	QNE	social	C		x				1.75
Weinstock et al.	2012	USA	unpub	535	CRT	general	G		x			x	14.00
Wood	1997	USA	unpub	731	OGPP	social	M	x					n.e.i.

Note. unpub=unpublished; pub=published; sample=sample of teacher; RCT=randomized controlled trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center; QWM=quasi-experimental design with matching procedure/equivalent groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design ; OGPP=one group pre-post measure; MBL=multiple baseline, social=training focus on social skills(e.g., behavior management); curriculum=curriculum implementation; quality=quality improvement; general=training focuses on learning & developmental appropriate practice; BM=behavior management; C=curriculum, Q=quality improvement; M=pedagogical /instructional model; SM=stress management, S=behavioral, interactional or relationship-based strategies; CCC=clinical consultation; G=general knowledge on developmentally appropriate practice *=random assignment of school district; n.e.i.=not defined.

2.4.1. In-service training impacts on social abilities

The findings on professional development effects on preschool teachers seem to be inconsistent. Some studies report positive effects on prosocial skills in young children (Baker-Henningham et al., 2009; Cordell, 2010; Domitrovich et al., 2007; Everston & Smithey, 2000; Girard et al., 2011; Landry et al., 2009; Rhodes & Hennessey, 2000; Sperry, 1999; Stoiber & Gettinger, 2011; Thanasetkorn, 2009; Vestal & Jones, 2004; Wood, 1997). Several other studies did not find in-service effects (Bryant et al., 2009; Driscoll et al., 2011; Graul & Zeece, 1990; Snyder et al., 2011). In addition, inconsistencies within studies related to outcome construct and informant appeared (Baker-Henningham et al., 2012; Gallagher, Abbott-Shim, & Vande-Wiele, 2011; Llyod & Millenky, 2011).

In-service delivery format is a potential explanation for the inconsistencies. Comparing approaches using *workshops and onsite* support, 6 studies demonstrated the impact of social-emotional skills training (Baker-Henningham et al., 2009; Domitrovich et al., 2007; Everston & Smithey, 2000; Girard et al., 2011; Sperry, 1999; Stoiber & Gettinger, 2011; Thanasetkorn, 2009). However, in one study, in-service training effects on prosocial behavior seemed to be inconsequential (Baker-Henningham et al., 2012). The promotion of social skills through *on-site support* solely was evaluated by 2 studies. So far, the evidence base does not support the presumption that social-emotional development can be fostered through quality improvement initiatives using coaching or mentoring (Bryant et al., 2009; Gallagher et al., 2011). The impact of *courses* as the main delivery approach has only been evaluated by 3 studies using social outcomes as the dependent variable. To date, the indirect influence of courses on social-emotional abilities has been demonstrated by 2 evaluations (Rhodes & Hennessey, 2000; Vestal & Jones, 2004) out of 3 studies (Graul & Zeece, 1990). The empirical foundation for the relationships between a *combination of workshop, course and onsite support* and social skills is vague. Only the study by Landry et al. (2009) suggested a positive impact on social skills when teachers implemented a literacy curriculum within the CIRCLE in-service program. Further, the impact of a *combination of course and onsite* support on developmental outcomes appears to be unclear. Only Cordell (2010) found an effect following the implementation of a literacy curriculum on the social skills of young children.

Assertions on the effectiveness of other in-service training delivery types are grounded only on single studies. Therefore, conclusions should be drawn with caution. Delivery formats using a series of workshops (Wood, 1997), and workshops in combination with ongoing on-site support and community of learner meetings suggest a relationship between these formats and positive child outcomes. For now, the efficacy of approaches using online resources solely or with consultation cannot be statistically verified (Driscoll et al., 2011).

Reviewing the results of in-service training effects concerning delivery format did not reveal clarification. Maybe *professional development content* provides a better explanation. The implementation of *specific behavioral strategies* in early childhood classrooms has been evaluated by 5 studies with 6 treatments. The findings from 4 treatments showed that social-emotional abilities can be improved through in-service programs that convey specific behavioral strategies (Driscoll et al., 2011; Girard et al., 2011; Sperry, 1999; Thanasetkorn, 2009). Further, professional development has been used to *introduce new curricula* into classrooms. However, training programs or curricula need to be implemented with high fidelity to have effects on child development (Arkoosh et al., 2007; Joyce & Showers 2002; Sterling-Turner, Watson, & Moore, 2002). The investigations by Cordell (2010) and Vestal and Jones (2004) conclude that professional development was effective in implementing curricula that lead to positive outcomes on social-emotional development.

Baker-Henningham et al. (2009) established a *behavior management system* taken from the Incredible Years program and the Dinosaur curriculum in Jamaican classrooms and found positive effects on the social development of young children. The establishment of behavior management through professional development has been evaluated extensively. Most of the studies investigated the implementation of the Incredible Years program or adapted versions of it (Baker-Henningham et al., 2009; 2012; Llyod & Millenky, 2011; Morris et al., 2010; Shernoff & Kratochwill, 2007; Snyder et al., 2011). Within the scope of behavior management programs, teachers integrated individualized and culturally sensitive interventions in the classroom. Behavioral management strategies should decrease classroom aggression and non-cooperation in classrooms, foster children's prosocial behavior and further school readiness. Within the group of studies which analyzed different Incredible Years implementations, no valid evidence on the impact of the introduction of a single behavior management intervention in classrooms (Baker-Henningham et al., 2012; Snyder et al., 2011) or the combination with stress management (Llyod & Millenky, 2011; Morris et al., 2010) on the social skills of young children was found. Positive effects on social-emotional outcomes in classrooms were seen in 2 studies implementing behavior management strategies other than the Incredible Years approach (Everston & Smithey, 2000; Stoiber & Gettinger, 2011).

So far, 3 studies have evaluated the impact of the implementation of a *multi-component model* to foster positive, and reduce negative, behavior in young children in early education settings. Baker-Henningham et al. (2009) observed a positive impact on appropriate behavior in classrooms after implementing the Incredible Years behavioral management system and the Dinosaur curriculum.

Other contents such as quality improvement, children's play, and literacy as well as developmentally appropriate practice were targeted by other professional development programs. The study by Landry et al. (2009) reported that social skills could be promoted through high-

quality literacy instruction and an enriched environment. Further, an evaluation of the IPPA Foundation course (Rhodes & Hennessey, 2000) also demonstrated improvements in the social play of children in center-based care in the UK. However, no significant improvements could be found for the quality improvement program examined by Bryant et al. (2009) or the Infant/Toddler Care Program (PITC) investigated by Weinstock et al. (2012).

2.4.2. In-service training impacts on reduced problem behavior

Concerning the reduction of inappropriate behavior, only 6 studies found a consistent effect of professional development on behavioral aspects. Certainly, 8 studies could not find any effects of teacher training. In 4 investigations, an impact was not obvious. Aggregating the findings systematically, it is possible to draw deeper conclusions on the influence of delivery format and specific content of programs that should be implemented in the classroom.

Eight studies analyzed the reduction of problem behavior through *workshop and onsite* in-service programs. Only one study was able to verify the relationship statistically (Filcheck et al., 2004), while in 6 investigations, no or discrepant effects were found (Baker-Henningham et al., 2012; Domitrovich et al., 2007; Everston & Smithey, 2000; Girard et al., 2011; Sperry, 1999; Stoiber & Gettinger, 2011). However, the reduction of challenging behavior in young children could be confirmed by 3 out of 4 studies measuring the impact of *onsite support* for preschool teachers (Demchak et al., 1992; Jo Rodriguez et al., 2009; Shernoff & Kratochwill, 2007). Only the implementation of the quality improvement project using the PFI Consultation Model (Bryant et al., 2009) did not demonstrate an effect on onsite and problem behavior. The studies evaluating the Foundation of Learning program did not support the assumption that appropriate behavior can be fostered through this professional development delivery type (Llyod & Millenky, 2011; Morris et al., 2010). Considering the small number of participants, the single-subject study by Hemmeter et al. (2011) provided a first insight that *data-driven feedback sent via email* offers the potential to improve teaching strategies and diminish challenging classroom behavior. Integrated child care centers in the Outreach Project were given the opportunity to request *individual workshops* based on their needs (Wood, 1997). The findings suggested that the developmental gap in months in social, emotional and behavioral aspects of children with disabilities declined after teachers took part in a series of workshops.

In line with the findings on social abilities, delivery format did not explain the heterogeneity of the findings. Therefore, the findings were aggregated in relation to the *content of professional development*. Only the study by Hemmeter et al. (2011) confirmed an in-service training effect on reduced problem behavior when *behavioral strategies* were implemented. Barnett et

al. (2008) reported that curriculum implementation through professional development resulted in a reduction of inappropriate behavior. Reducing behavioral problems indirectly through professional development, forcing the establishment of behavioral management, seems to be possible (Demchak et al., 1992; Jo Rodriguez et al., 2009). The study by Shernoff and Kratochwill (2007) supported the assumption that problem behavior can be reduced when a behavior management system is implemented through in-service training. The findings from Baker-Henningham et al. (2009) also indicated an in-service training effect after implementing behavior management techniques in combination with the Dinosaur Curriculum. Inconsistent or no effects were found in 3 studies implementing *behavior management strategies* only (Baker-Henningham et al., 2012) or combined with stress management (Llyod & Milenky, 2011; Morris et al., 2010). A *multi-component model* was used by Filcheck et al. (2004). Teachers received support to implement a multi-level intervention; they also received additional training to implement child-direct interactions in classrooms. Further, a parent-child interaction component was added to the professional development. The findings showed a decrease in inappropriate behavior in children in the classroom after new contents and training were added (ibid.).

2.4.3. Challenges measuring in-service training effects on social skills and problem behavior

The aggregation of findings suggests that it is possible to foster social-emotional skills through professional development. In-service training was less effective in reducing problem behavior. On the one hand, more intensive efforts might be needed to decrease problem behavior. At this stage, conclusions cannot be drawn with confidence whether more *intensive in-service* training approaches are needed or whether other or more intensive instructional strategies (i.e., child-centered approaches vs. whole-class interventions) are needed to decrease challenging behavior. On the other hand, *outcome construct* is known to influence the effectiveness of interventions (Wilson & Lipsey, 2001). According to the systematic review, it can be assumed that in-service training effects are larger for social skills than for the reduction of problem behavior. This could be empirically examined through a meta-analysis. Further, meta-regression would provide further insights if effects were moderated by the intensity of training or content. *Training intensity and duration* are widely acknowledged as potential effect modifiers, but have not been systematically examined yet.

The results also point out that professional development might be moderated by *delivery format and content*. However, their influence on the magnitude of in-service training effects is not clear so far. Inconsistent effects following the introduction of the same content indicate

that the delivery of in-service programs might have more influence on the successful implementation of new practices in classrooms than the content. This consideration is underpinned by the fact that the effectiveness of the Incredible Years Program has been empirically validated many times (Incredibly Years, n.d.), but when different training formats were used to implement it, their effectiveness was divergent. Results show that some professional development contents are more effective than others (e.g., curricula vs. behavior management). Differential effects would be expected for in-service programs that focus on social development, language and literacy, and quality improvement. Comparing professional focus and content through a meta-analysis would be the next step.

Further, reported effects are based on descriptive statistics or statistical procedures like analysis of variance or regression with dummy codes. Effect sizes are barely reported, even though this is recommended by the American Psychology Association (American Psychology Association, 2009). The reported studies were diverse with respect to study design, methodological quality, and sample size. Variations in studies need to be analyzed with meta-analytic techniques to systematically examine the roles of moderators.

To sum up, the systematic reviews provide preliminary evidence that it is possible to improve quality in early childhood education and to enhance student achievement. It is still uncertain how large the impact of in-service training is concerning the magnitude of effect sizes for a specific treatment. The reviews point out that the results from primary studies on in-service training impacts are heterogeneous. In line with the theoretical concept of the integrated framework, some potential explanations of in-service effectiveness have emerged in the literature syntheses. In particular, delivery format (how) and content of training (what) have been discussed as potential effect modifiers. At this stage, some delivery formats seem to be more effective than others, but it is not possible to identify the most promising delivery type through a systematic literature procedure. This conclusion can be repeated for the content and focus of in-service training. In addition, differential effects of in-service training were found for different groups of recipients (who), including mono- and bilingual children as well as typical learners and children at risk. Further, primary studies that evaluated in-service effectiveness were diverse in terms of study design, measurement features, and method of analysis. Converting the results to effect sizes would make it possible to combine and compare the findings through meta-analytic techniques. Meta-regression and subgroup analysis would be the next step to explain the heterogeneity of in-service training effects empirically.

3. Methodology

3.1. Research aims and problem formulation

The primary research aim of this set of meta-analyses was to aggregate the evidence on the impact of in-service professional development programs for early childhood educators on external quality ratings and child development in the areas of literacy, language, social-emotional and behavioral skills. Further, the project also aimed to investigate and explain the heterogeneous results of in-service training effectiveness. In-service professional development is defined as training opportunities for early education professionals working in center-based care, serving children from birth to entry in formal schooling. International comparisons clearly indicate that there are several in-service training opportunities, differing with respect to content, format and provider (OECD, n.d.). The existing formats (e.g., workshop, course, onsite, community of practice or online) are therefore, included in the definition of in-service programs. The impact of in-service professional development has been measured through randomized-experimental, quasi-experimental and non-experimental studies. The outcomes of in-service professional development within this set of meta-analyses are a) external quality ratings made through systematic observations and/or b) child development. Child development is measured through assessments, observations, and mothers' or teachers' ratings.

3.1.1. Research questions

The meta-analyses investigated the in-service training effectiveness on quality ratings and child development. Evidence-based decision making relies on research syntheses that guide practitioners and policy makers. Recently, new emphasis has been placed on how research is conducted, what the findings are and what the cumulative evidence suggests (American Psychology Association's Presidential Task Force on Evidence Based Practice, 2006). Several factors regarding methodology and study quality are known to explain between-study and within-study variance (for an overview see Higgins & Green, 2011; Wilson & Lipsey, 2001). Before examining the influence of intervention features, methodological aspects were tested within every meta-analysis. The methodological aspects at study level incorporated: a) randomization and assignment of groups, b) attrition rate, and c) sample size (at beginning). At the effect size level, the following methodological issues were considered: a) informant, b) contrast or change, c) pretest differences, d) score level (e.g., total score, subscale or single

item), e) standardized/established or newly developed instruments, f) reliability scores (i.e., scale reliability with $\alpha > .70$ and inter-rater reliability with ICC or *Kappa* $> .80$). Further, publication type and publication year were coded and examined as potential moderators. Aspects concerning publication bias and study characteristics were addressed first, before theoretical modifiers were examined.

Several research questions emerged after conducting the extended literature search and review. Literature reviews show that the evidence base and theoretical models vary, depending on outcome domain. Therefore, separate meta-analyses with separate data sets were conducted. Exploratory questions were examined separately for each set of data (see table 8).

Table 8. Research questions according to outcome domain

Meta-analysis: In-service training impacts on external quality ratings

- To what extent can the quality of early childhood settings be improved through in-service training programs?
- Are there different effect sizes according to different quality constructs and scales?
- Does the implementation of a curriculum moderate the effectiveness of in-service training approaches?
- Do training intensity and duration influence the effectiveness of in-service training?
- Are there significant differences in the effect sizes between different professional development methods (e.g., workshop, course, onsite etc.)?
- Are in-service programs with coaching more effective than other approaches?
- When professional development is brought to scale, is it more or less effective?
- Are scale-based in-service training approaches more effective than others?
- Does the workplace environment (e.g., provider, setting, classroom composition) influence the in-service training effect?

Meta-analysis: In-service training impacts on language and literacy abilities

- To what extent can language and literacy abilities of young children in center-based care be affected by in-service professional development?
- Do in-service training programs affect language and literacy abilities equally?
- Do training intensity and duration influence the effectiveness of in-service training?
- Do type and delivery formats of in-service programs moderate their effectiveness?
- Do children from different backgrounds benefit equally from teacher training?
- Is the fidelity of implemented strategies a significant predictor of the training effect?

Meta-analysis: In-service training impacts on social-emotional skills and behavioral competencies

- To what extent can in-service professional development foster social abilities?
- To what extent does in-service training reduce challenging behavior?
- Do training intensity and duration influence training effects on social and behavioral skills?
- Do type and delivery formats of in-service programs moderate their effectiveness?

Meta-analysis: Impact of changes in early childhood education quality on child development

- To what extent do quality improvements during training predict in-service training effects on child development at posttest?
 - How much of the variance in in-service training effects on child development can be explained by changes in child care quality during training?
-

3.1.2. Selection criteria

Selection criteria were defined to investigate in-service training effects. The selection of relevant studies was guided by the following criteria:

1. Studies must focus on in-service professional development for early childhood practitioners working in center-based child care.
2. Studies must use quantitative outcome measures related to child care quality and/or child development.
3. Study sample must contain children aged zero to seven in center-based care.
4. Studies must use either experimental, quasi-experimental or one-group pre-post designs.
5. Studies must report sufficient statistical information to compute effect sizes.
6. Studies must be published in English, German or French.

A broad definition of in-service professional development, including a variety of training formats, was used for 2 reasons. First, international comparisons show that continuous and ongoing professional development provided to early childhood professionals is diverse (OECD, n.d.). Second, the broad definition allows for the examination of variations in format as potential moderators of training effectiveness. As mentioned earlier, the early childhood workforce constitutes a group that is diverse with respect to their qualifications, experience, roles (lead teacher, assistant), organizational affiliations (centers, preschools, kindergartens, public school programs) as well as personal and cultural characteristics (NPDCI, 2008). The definition “center-based care” covers a diversity of workforce and settings in early childhood education, but clearly distinguishes between early education in centers and family day care. Studies measuring the impact of professional development on center-based care and family day care were included when outcomes were reported separately for center-based care. In addition to the great diversity of the child care workforce, children from various backgrounds are enrolled in regular center-based care and inclusive approaches and classrooms are established in early childhood education. Inclusive classrooms or children with diagnosed developmental delays in early childhood centers were included in the meta-analyses. However, services which targeted a specific group of children (e.g., at risk, immigrants, disadvantaged) with the in-service intervention were coded as a potential moderator.

A critical concern of meta-analyses is to choose the right research design to evaluate the impact of an intervention or training. The method of meta-analysis has often been criticized (“mixing apples and oranges”) due to the fact that different research forms (e.g., randomized controlled trials [RCT] or non-randomized quasi-experimental studies) are included to investigate a specific phenomenon. It has often been argued that investigations of impacts should be limited to randomized controlled trials, because effect size is affected by confounding var-

iables (Borenstein et al., 2009; WWC, 2011; 2013). Several risks of bias have been discussed regarding the mechanism of allocation (Higgins & Altman, 2008). Often selection bias is assumed when randomized assignment is missing, indicating that participants are invited instead of required to take part in the intervention (Borenstein et al., 2009). In these cases, a motivation bias is presumed, in which the motivation of participants influences the outcome. In particular, in the discussion of American standards favoring randomized controlled trials (Mosteller & Boruch, 2002; National Research Council, 2004b; 2005) recommendations for systematic reviews and meta-analysis have been made to analyze RCT and quasi-experimental studies separately (Borenstein et al., 2009). Seidel and Shavelson (2007) argue in their meta-analysis that RCT and quasi-experimental studies should be combined when the evidence base is limited. Existing reviews indicate that the number of experimental studies is limited (Fukkink & Lont, 2007; Zaslow et al., 2010b). According to the recommendation by Hall, Rosenthal, Tickle-Degnen, and Mosteller (1994), it is appropriate to aggregate studies with different formats when they measure the same phenomena. They suggest that the triangulation of methodologically different studies lends reliability to the validity of an effect. Aspects of motivation biases are listed as well as the separation of experimental and quasi-experimental studies. Recent research points out that only a marginal number of early child care teachers (< 5 %) do not participate in ongoing professional development (Behr & Walter, 2012; Fuchs-Rechlin, 2007). The findings indicate that motivational aspects do not affect whether teachers participate in in-service programs or not, but it might pertain to what kind of in-service program they want to receive. Further, international reports demonstrate that ongoing professional development for the early childhood workforce is considered as a duty and compulsory in several European countries (Eurydice, 2014). Regarding current early childhood circumstances and the limited number of randomized controlled trials, experimental and quasi-experimental studies were initially combined in the meta-analysis.

3.2. Literature Search

Relevant studies, published between 1970 and 2011, were retrieved through a systematic literature search. First, an electronic search in English-speaking and German-speaking data bases was conducted. Second, studies were individually searched for in bibliographies, review articles and renowned journals. Additionally, conference and convention programs were systematically screened.

3.2.1. Electronic search

An electronic search was conducted in ERIC, PsycINFO, ProQuest Dissertations and Theses, ProQuest Educational Journals, WISO (Literatur für die Wirtschafts- und Sozialwissenschaften), and the FIS Bildung Literaturdatenbank. The systematic electronic search for the meta-analysis was limited to literature published between 1970 and 2011. The electronic data bases provide a huge variety of studies and mixture of published and unpublished research, such as journal articles, conference papers, dissertations, technical and governmental papers, books and chapters. Table 9 lists the key words used in the electronic search.

Table 9. Keywords for electronic search

Descriptor	Keyword or Thesaurus (English)	Keyword (German)
In-service professional development	"in-service" or "professional development" or "coaching" or "consulting" or "technical assistance" or "mentoring" "teacher training" or "teacher education"	"Weiterbildung" oder "Fortbildung"
Outcome measure	"quality" or "performance" or "teacher behavior" or "child outcomes" or "program improvement" or "enrichment" or "environment"	"Performanz" oder "Entwicklung" oder "Qualität"
Target	"impact" or "effect*" or "effectiveness"	"Effekt" oder "Effektivität" oder "Wirkung"
Education level	"early childhood education" or "kindergarten" or "preschool" or "child care"	"Erzieherin"

3.2.2. Hand search

A hand search was conducted to complement the systematic electronic search. Studies were searched for in previously published systematic reviews and meta-analyses (i.e., in Fukkink & Lont, 2007; Klein & Gomby, 2008; Zaslow et al. 2010b). Further, the reference list of studies that were identified as relevant for meta-analysis was also reviewed to identify more literature that is relevant. Additionally, the following renowned journals were screened: Early Childhood Research Quarterly, Journal of Early Teacher Education, Journal of Teaching and Teacher Education, and Frühe Bildung.

Professional meetings and conferences play an important role in the dissemination of scientific information about newly completed research (Garvey & Griffith, 1979; Rosenthal, 1994). For this reason, attendance at professional meetings and searching in the conference programs of renowned meetings are appropriate ways to locate at least name, title and abstract of diffuse information that focuses on the relevant topic. The publication process for scientific papers is a slow one, so sending a written request to authors for copies of their presentations is necessary to obtain newly completed research (Rosenthal, 1994). To gather the latest re-

sults, conference proceedings and venues were screened at least for the last 3 conferences years (2009 - 2011).

- Biennial Meeting of the European Association for Learning and Instruction (EARLI)
- Biennial Meeting of the EARLI Special Interest Group 5
- Kongress der Deutschen Gesellschaft für Erziehungswissenschaften (DGfE)
- Conferences of the Society for Research on Educational Effectiveness (SREE)

A web-search through Google was also carried out. The keywords used in the electronic search were used. Further, combinations of key word were utilized (i.e., “impact of in-service professional programs for preschool teachers”). Apart from the identification of one German study after a systematic search, the amount of relevant literature in German was unrewarding. Therefore, an extensive Google search of German literature was made in a subsequent phase. The following Google search terms were used: “Wirkung von Weiterbildungen für Erzieherinnen”, “Effekte von Erzieherinnenweiterbildung“, “Effekte von Coaching für Erzieherinnen”, „Weiterbildung für pädagogische Fachkräfte“.

3.2.3. Search Process

The literature search in electronic data bases resulted in an initial set of 1,162 hits (see table 10). The initial set of hits was narrowed according to the following criteria: (1) investigations of in-service programs were designed to improve quality of child care or child development, (2) sample included preschool, pre-k, or kindergarten teachers as well as educators working in center-based care, (3) studies used a quantitative research design (control group comparison or one group pre-post measure) and (4) studies were published in English, German or French. After the title and abstract screening by 2 independent coders, 202 hits were identified as being relevant for the set of meta-analyses on professional development impacts on quality ratings and child outcomes. Listed in table 10, several studies were found in international data bases, but an unsatisfactory yield was gained from German data bases (WISO and FIS Bildung). In addition to the electronic search, journals, conference venues and reference lists of relevant studies were screened to find supplemental literature. A search through the web-engine Google was made for English papers but even more extensively to gain German literature. The web-search resulted in 8 additional German papers. Overall, 235 references were identified as relevant and ordered or downloaded, but only 231 were available⁵.

⁵ The following 4 references were not available: Askin, B. E. (1978). Responsive Environment Early Education Program (REEEP) Third-year evaluation study. Final evaluation report, 1977-78.; Lopez, K. (1991). Improving an extended day care environment through staff training and activity development. Unpublished dissertation. Nova University; Spillman, S. D. (2011). A comparative study of two professional development models' impact on pre-

A full text review procedure (Screening II) with a short coding form was used to evaluate the quality of the 231 studies. In addition to title and abstract screening, further inclusion criteria were: (1) measures of quality in early childhood classrooms through observation or measures of child development, (2) explicit in-service training for teachers/educators serving children from birth to formal schooling, and (3) sufficient statistical information to calculate effect sizes. From the 231 studies, 85 were identified as relevant and as providing sufficient statistical information (see figure 5). Further, 44 papers were also categorized as relevant, but did not provide sufficient statistical data. The main authors of the 44 papers were contacted via email 3 times. Twenty-one authors responded and of them, 10 authors provided the missing statistical information for 10 studies.

Table 10. Literature search

Data base	Hits	Relevant for analysis
Electronic Search		
ERIC	703	124
PsycINFO (via OVID)	77	25
ProQuest Dissertation & Thesis	84	20
ProQuest Education Journals	151	32
FIS Bildung	126	0
WISO	21	1
Results electronic search	1162	202
Hand search		
Journals	-	3
Reference lists	-	18
Conference venues	-	4
Web-based search engines (google)	-	8
Identified to be relevant	-	235
Available for coding	-	231

In summary, the multi-step literature search procured a total of 95 studies that met all criteria and were included in the meta-analysis (see figure 5). Of these, 36 studies with 42 different professional development treatments were found that measured the impact of training on quality ratings. Further, 59 studies were found that measured the impact of in-service programs on child development. Of the 59 studies, 48 studies evaluated the impact of 59 different training approaches on language and literacy scores and 13 studies with 14 treatments provided information on social and behavioral aspects of development. Of the mentioned evaluations, 9 studies investigating 10 different treatments were found that considered both data on quality ratings and child development.

Six additional studies were found with child outcomes. Four studies measured training effects on math development, one study on real life problem solving and one study on executive functions and motor development. A potential problem in meta-analytic techniques occurs

when the aggregation of findings and statistical synthesis are based on a small number of studies. In particular, serious challenges appear when the variance between studies is estimated. The standard error of the aggregated effect in random effect models is partially based on the estimation of the variance between studies and estimations using an insufficient number of studies are substantially defective (Borenstein et al., 2009). At this point, conducting a meta-analysis on the impact of in-service training on math abilities in young children does not make sense, as there are only a limited number of studies ($n = 4$).

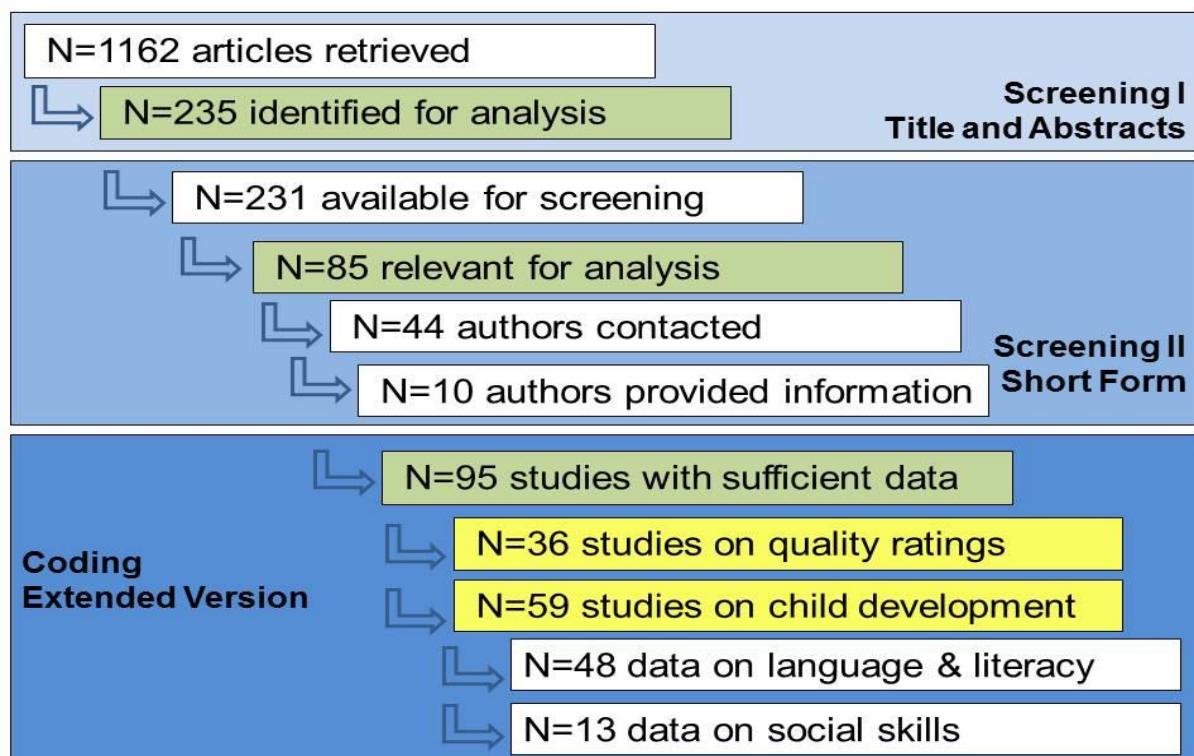


Figure 5. Literature search process

3.3. Coding and gathering information from studies

A multi-step full-text coding procedure was used to systematize the information from the studies. Two independent reviewers coded all the studies. First, the studies identified as being relevant from the title and abstract screening were coded with a short screening form to evaluate their quality and to assess if there was sufficient statistical data. The short screening form included coding on study design (e.g., control-group design, randomization, sample size), instructional features of the professional development, adequacy of the statistical information and relevance for meta-analysis (see appendix A). Inter-rater reliability was excellent for study relevance ($Kappa = 1.0$).

Second, relevant studies were coded with an extended coding schema (see appendix B). The coding schema gathered information on the independent variables describing study characteristics, in-service interventions, and context. Through a formatted excel sheet, statistical data were gathered to calculate effect sizes (dependent variable). In the excel sheet, relevant information on other independent variables (e.g., scales and instruments, scoring levels, reliability of scales, informants and type of contrast) were coded at effect size level. Based on the guidelines from the Campbell Collaboration⁶ (n.d.), all studies were double-coded by 2 independent coders and disagreements between the coders were resolved through discussion and consensus. In the discussions, both coders reviewed the papers simultaneously and made a final decision. Four evaluators carried out the coding procedure. Inter-rater reliability for nominal variables was estimated with *Kappa* and for ordinal and interval scaled variables with Intra-Class-Correlations⁷. Only reliable codes were included in the meta-analyses. The coding variables and inter-rater reliability are reported in the next section.

3.3.1. Coding at study and treatment level

The studies were coded extensively with regard to study descriptors and intervention features at study or treatment level, based on the recommendations by Wilson and Lipsey (2000). An extended coding schema was used for every treatment because some studies reported multiple experimental comparisons (e.g., Buschmann et al., 2010b) or multiple treatments (Assel et al., 2007; Japel, 2009; Lonigan et al., 2011; Neuman & Wright, 2010) within a paper. Information on publication type (1=published in journals or books/0=unpublished technical or governmental report, dissertation or conference proceedings) and publication year were coded with excellent inter-rater reliability⁸. Methodological aspects and descriptors for child care settings and interventions are illustrated in table 11. There was excellent inter-rater reliability (*ICC* >.92) for the variable “total sample at beginning of study” in both data sets. Attrition rates were not reported consistently in all the studies and the coders estimated the attrition rates in percentages using information on the samples at the beginning and the number of participants at posttest analysis. High agreements were found for

⁶ The Campbell Collaboration C2 is a non-profit organization that maintains and disseminates systematic reviews in the fields of education, crime and justice, and social welfare.

⁷ Agreement of the 4 coders (at single measure) was estimated through a two way random model with absolute agreement to examine systematic errors regarding measure and person.

⁸ Several recommendations exist for the interpretation of Cohen's *Kappa* (Greve & Wentura, 1997). Landis and Koch (1977) define Cohen's *Kappa* with a value of >.40 as moderate, >.60 as substantial, and >.80 as almost perfect inter-rater reliability. Bakeman and Gottman consider a *Kappa* of <.70 as unacceptable. Wirtz and Casper (2002) suggest that an *ICC* score >.70 indicates good reliability (p. 160).

attrition rates of children ($ICC = .87$), but discrepancies appeared for the coding of teacher attrition rates ($ICC = .50$ and $.62$). Good inter-rater reliability estimates ($ICC > .70$) were found for comparison type (e.g., placebo, business as usual, or alternative treatment). Coding on the assignment of teachers to intervention and control groups was less definite ($ICC = .38$ to $.63$), because of vague reporting in the research literature. Excellent agreement was found for coding on the variable “percent of female children in intervention group”, indicating unbalanced grouping.

In the literature, the characteristics of settings and participants were considered to influence in-service training effectiveness. Satisfactory inter-rater reliability was found for setting variables and teacher characteristics, at least for the investigation of child outcomes. Classroom characteristics were coded to identify whether the majority of children were classified as being at risk and whether the children in the study were in the age group of under 3 years of age or not. A dichotomous variable for child care providers was used, separating public programs for disadvantaged children (e.g., Head Start, Perry Preschool) and other programs. Inter-rater reliability was good for the investigation of child outcomes, but unsatisfactory for the data set on quality ratings. The coding of moderators concerning teachers’ educational background (presence of a university degree) revealed high reliability scores.

Concerning the content and topic of professional development programs, the coding was vague due to imprecise reporting in studies. For the subset of studies examining in-service training effects on child outcomes, most studies focused on multiple aspects of professional development. Therefore, it was not always possible to distinguish what the predominant contents and topics of the in-service program were⁹.

Categorizing professional development into 5 different delivery methods (i.e., workshop, course, online learning, onsite and community of practice), moderate to almost perfect inter-rater reliability was calculated for delivery types ($Kappa = .52$ to $.87$), except for community of practice. Distinguishing between different onsite formats, substantial reliability was calculated for consultation and mentoring. For the coaching category, the raters did very well for the subset of studies on child outcomes ($Kappa = .80$). For the subset of studies on quality ratings, coding was vague, with 3 disagreements out of 42 studies ($Kappa = .48$). Ratings, whether only onsite or mixed delivery methods were used, obtained satisfactory agreement scores. The question of whether the professional development was a scale-based program was answered reliably for the subset of studies on quality ratings.

⁹ The topics of in-service programs should distinguished between the following options: 1) first aid & health, 2) math and science, 3) language and literacy, 4) social-emotional and behavioral development, 5) nutrition and sport, 6) quality improvement, 7) implementation of curriculum or model, 8) other topic.

Table 11. Inter-rater reliability of study and treatment level

Moderator	Quality		Child outcomes	
	<i>Kappa</i>	<i>ICC</i>	<i>Kappa</i>	<i>ICC</i>
Publication				
Publication year		.98		.98
Publication type	1		1	
Methodological characteristics				
N total sample of teachers		.97		.93
N total sample of children		.93		.95
Attrition rate of teachers in %		.50		.62
Attrition rate of children in %		-		.87
Comparison type (4 options given)		.72		.88
Teacher assignment to treatment (6 options given)		.38		.63
% of female children in intervention group		-		.96
Characteristic of setting				
Majority of children at risk (1=yes /0=no)	.69		.60	
Children under 3 years of age in sample (1=yes/0=no)	.64		.83	
Context of study (1=Head Start/0=other child care)	.37		.72	
All teachers in sample have a university degree (1=yes/0=no)	.63		.87	
% of teachers with university degree in intervention group		.99		.81
Content of in-service program				
Curriculum implementation major goal (1=yes/0=no)	.70		.10	
Major topic (8 options given)		.57		.15
Delivery format				
Workshop (1=yes/0=no)	.62		.87	
Course (1=yes/0=no)	.52		.79	
Online training/blended learning (1=yes/0=no)	1		.82	
Community of practice (1=yes/0=no)	.21		.30	
Onsite support (1=yes/0=no)	.75		.73	
Coaching (1=yes/0=no)	.48		.80	
Consultation (1=yes/0=no)	.65		.71	
Mentoring (1=yes/0=no)	.91		1	
Only onsite support used (1=yes/0=no)	.73		.88	
Mixed formats used in in-service program (1=yes/0=no)	.63		.77	
In-service training is scale-based (1=yes/0=no)	.57		.12	
In-service training intensity				
In-service training duration (in months)		.82		.90
In-service training amount (in hours)		.78		.86
In-service training longer than one year (1=yes/0=no)	.40		.85	
In-service training longer than 30 hours(1=yes/0=no)	.24		.87	
Confounders in in-service programs				
N of in-service training providers		.39		.73
Person delivering the in-service program (6 options)		.44		.71
Fidelity of curriculum implemented	-		1	

Note. Comparison type (1=business as usual/2=placebo/3=different treatment levels/4=no control group); teacher assignment (1=randomization at individual level/2=randomization at classroom level/3=randomization at center level/4=matching individuals/5=matching classrooms/centers/6=groups of convenience); delivering person (1=researcher/2=graduate student/3=professor/4=professional coach or tainer/5=multiplier/6=specialist or therapist).

Concerning training intensity, very reliable categorizations¹⁰ were made for duration in months and amount of in-service training in hours ($ICC > .77$). For in-service training hours, summed scores of all training hours for different delivery types were calculated¹¹. Training intensity was coded in dichotomous categories; these showed high agreement for studies on child development, but very low agreement for the subset of investigations on quality ratings. It was hypothesized that the quality of professional development might be confounded by the person who delivers the training and by the number of facilitators needed. From the initial short screening, 6 types of delivery persons were found to be usual: 1) the researcher, 2) graduate students, 3) a professor and faculty member of the university or college, 4) a professional coach, consultant or trainer, 5) a multiplier who received training earlier, 6) specialists such as speech and language pathologist/therapists. However, not all studies reported how many trainers or coaches and what kind of trainer were recruited for the study. While substantial reliabilities between the coders in the categories of number of providers and delivery persons were found for the subset of studies on child outcomes, less good agreements appeared for the studies on quality ratings.

The literature suggests that in-service training is more effective when specific curricula are implemented with high fidelity (cf. Klein & Gomby, 2008; Justice et al., 2008; Jo Rodriguez et al., 2009), however, only 3 studies reported fidelity indices. The inter-rater reliability ($ICC = 1$) was perfect in that category, but the number of studies was too small to make solid comparisons.

Further, it was initially intended to code and analyze variables on transfer factors in in-service interventions and training in more detail, but insufficient reporting of such features prevented their inclusion for further analysis (see table 12). In particular, trainee characteristics, the workplace environment, and training design were mentioned as significantly moderating training transfer and effectiveness. The available information on trainees consisted of years of experience, age of teachers, their role in the center (e.g., lead teacher, assistant, or director) and other information on participants. Further, a supportive workplace environment was coded dichotomously depending on whether the whole team or the director, too, participated in the in-service program. It was intended to extract facts on specific components of the training (e.g., training on assessment, working with parents, classroom management, etc.), and the number and rhythm of sessions, but this information was only available for a limited number of studies. Specific professional development formats such as scale-based training and community of practice could not be coded with tolerable reliability.

¹⁰ Categories were made referring to student credit hours. One credit hour is equivalent to approximately 15 contact hours per semester. A semester is equal to 6 months and half of a semester is equal to three months.

¹¹ When a summed score on all training hours was not reported, the coder calculated a summed score. For example, when a full training day was reported, the coder assumed 8 hours and for a half day, 4 hours.

The quality of in-service programs, in particular the presence of a training manual and fidelity checks, could not be included in the meta-analysis due to insufficient reporting. The pre-knowledge, pre-training and motivation of participants were not reported in the primary studies. Additionally, more detail was aimed for coding the settings in which the studies were conducted (e.g., classrooms serving disadvantaged children, age-homogeneity in classrooms, inclusive classrooms with disabled children, socioeconomic index for classrooms), but reporting was vague and resulted in low inter-rater reliability scores. In the coding procedure, 3 categories of data for potential effect modifiers were obtained (see table 12).

The first category of moderators had high inter-rater agreement and met the standards of Cochrane¹² and Campbell Collaboration¹³. The second category did not meet the standards of Cochrane Collaboration for high inter-rater reliability (*ICC* or *Kappa* < .70), but it did meet the standards of Campbell Collaboration using a double coding procedure for all variables with a discussion of coding on disagreements. Therefore, the second category of moderator variables met the required standard with reservations and the results must be carefully interpreted. The third category of moderators is excluded from the meta-analysis, because research papers did not provide information on them.

Table 12. Inter-rater reliability of potential effect modifiers

Moderator variable met standards (<i>ICC</i> or <i>Kappa</i> > .70)	Moderator variable met standards with reservations (<i>ICC</i> or <i>Kappa</i> < .70)	Exclusion of moderator due to missing information
Publication year	Attrition rate of teachers	Teacher experience
Publication type	Teacher assignment to treatment	Teacher age
Sample size teachers	Person delivering in-service training	Teacher role
Sample size children	Majority of children at risk	Pre-knowledge and pre-training
Comparison type	Curriculum implementation	Motivation of participants
Attrition rate of children	Major topic	Team approach
% of female children in EG	Community of practice	Director participation
% of teachers with university degree	Children under 3 in sample*	Fidelity of in-service training
In-service training duration	Context of study*	Classroom: age homogeneity
In-service training amount	All teachers have university degree*	Classroom: inclusive setting
Mentoring	Workshop*	Classroom: ses index
Online training	Course*	Classroom: disadvantage rate
Onsite support	Coaching*	
Fidelity of curriculum	Consultation*	
	Mixture of training formats*	
	Scale-based training*	
	In-service training longer than one year*	
	Number of in-service providers*	

Note. *=variable met standards for the meta-analysis on child outcomes, but met standards with reservations for the meta-analysis on quality ratings.

¹² The Cochrane Collaboration is an international non-profit organization in health care and medicine that conducts and disseminates systematic reviews and meta-analyses. It is related to the World Health Organization.

¹³ The Campbell Collaboration is a non-profit organization for preparing and disseminating systematic reviews for evidence-based policy in social research including education, crime and justice and social welfare. It is a sister initiative of the Cochrane Collaboration.

3.3.2. Coding at effect size level

In addition to the ratings on study and treatment levels, coding of moderators at effect size level of extracted statistical data for meta-analysis, mainly methodological issues, enabling them to be contrasted. For this purpose, another coding strategy was used. First, the main reviewer extracted relevant statistical information (e.g., M, SD, N, effect direction) and entered it in an excel sheet and coded methodological information (e.g., reliability, score level, informant, outcome domain). Second, an additional reviewer verified and checked all the coding and highlighted discrepancies. Third, discrepancies were resolved through reviewing the paper simultaneously. Table 13 shows the number of discrepancies that were corrected.

Table 13. Inter-rater disagreement in percent in coding on effect size level

Coded variable	Quality k=475	Child outcomes k=729
	% of corrections	% of corrections
In-service training group (M)	4.63	2.04
In-service training group (SD)	4.42	2.77
In-service training group (N)	2.11	1.89
Control group (M)	3.79	2.04
Control group (SD)	4.21	2.47
Control group (N)	2.53	1.89
Effect direction (1=automatic/2=positive/3=negative)	0.42	3.93
Treatment comparison type (1=contrast t2/0=change t2-t1)	3.58	1.23
Pretest differences (0=no/1=EG > CG/2=CG > EG)	0.00	0.00
Standardized instrument (1=yes/0=no)	0.00	6.11
Score level (1=total/2=domain/3=subscale/4=item)	11.16	7.86
Inter-rater reliability (1=ICC or <i>Kappa</i> >.80/0=<.80)	0.00	8.30
Reliability (1= α >.70/0= α <.70)	0.00	8.88
Informant (1=child/2=teacher/3=observer/4=parents)	-	5.39
Quality scale (1=ERS/2=ELLCO/3=(in) CLASS/4=other*)	0.42	-

Note. EG=experimental group; CG=control group; *= ratings based on other observational scales (Early Childhood Classroom Observation Measure [ECCOM]; Early Childhood Classroom Observation Scale [ECCOS]; ECE Literacy Observation Checklist; Early Literacy Activity Rating Scale; Assessment Profile: Subscale Learning Environment; Supports for Early Literacy Assessment [SELA]; amount of instructional time).

Overall, very few corrections were needed for the extraction of statistical information (< 5 %). The reviewers had very high agreements (> 90 %), except for coding on the score level of quality ratings (11.16 %). No corrections were needed for the coding of the variables of standardized instrument, inter-rater reliability and reliability within the data set of quality outcomes. This is due to the facts that a) either a renowned quality scale (ERS, ELLCO, CLASS) or an unknown, self-developed instrument or checklist was used and b) information on the reliability of scales and inter-rater agreement statistics was usually either excellent or

missing. In particular, for one third of the effect sizes ($k = 190$) reliability scores were not reported and for more than half of the effects ($k = 271$) inter-rater reliability was not reported.

Table 14. Systematization of child outcomes for coding

Coding	Descriptors
Language skills	
Receptive vocabulary	Assessments involving receptive vocabulary (PPVT, TVIP, PRE CTOPPP receptive vocabulary, PLS complex receptive vocabulary)
Expressive vocabulary	Assessment and ratings involving expressive vocabulary and language (EOWPVT, IGD picture naming, expressive vocabulary tests, PRE CTOPPP expressive vocabulary, LELA expressive language, PLS expressive language, ELAN, SELDAK)
Utterances	Analysis of children's utterances from videos or transcripts involving children's responses (multi-word response, level of response, peer-directedness, story comprehension response) and linguistic categories (verbal productivity, lexical diversity, or syntactical complexity)
Language comprehension	Assessment involving language comprehension (WJ passive comprehension, PLS auditory comprehension, DSC auditory comprehension, SETK Verstehen von Sätzen)
Language composites	Assessment and ratings involving language scores (IDEA oral language proficiency test, MCDI vocabulary score, ECI Early communication indicator, WLPB language, Creative curriculum development continuum language, Beller Entwicklungstabellen Sprache, SELDAK, SSMIK, SETK, HSET)
Literacy skills	
Alphabet & letter knowledge	Assessment involving alphabet knowledge, naming letters and letter-word identification (PALS Pre-K alphabet, EOWPVT letter knowledge, ELS letter naming, LELA upper and lower class letter record, PRE CTOPPP letter sound & naming letters, WJ letter-word identification, naming letter task, CAP letter, DIBELS letter naming fluency)
Blending	Assessment involving blending abilities to determine words (PRE CTOPPP blending, ELS blending, blending task)
Elision	Assessment involving elision abilities including the omission of sounds, syllables or words (PRE CTOPPP elision)
Concepts about print	Assessment and ratings involving print knowledge and concepts about print (PRE CTOPPP print knowledge, print awareness, CAP concepts about print; FACES book knowledge, LELA book knowledge, TERA book handling, SSMIK & SELDAK Selbstständiger Umgang mit Büchern)
Rhyming	Assessment involving rhymes (ELS rhyming, IGD rhyming, alliteration, PAT rhyming, rhyming production)
Initial/beginning sound	Assessment involving beginning sound identification (ELS beginning sound, letter sounds, LELA beginning sound, WJ sound identification, DIBELS initial sound fluency)
Writing	Assessment and ratings involving (early) writing (ELS word writing, CAP writing, SELDAK Schreiben)
Phonemic awareness	Assessment involving phonemic awareness (PAT phonemic awareness task, LELA phonemic awareness, PALS phonemic awareness)
Phonological awareness	Assessment involving phonological awareness (CTB phonological awareness)
Segmentation	Assessment involving segmentation abilities (PAT onset segmentation, ELS phonemic segmentation, Yopp Singer total word segments, Segmentation task)

Note. Table 14 provides an overview of instruments included in the meta-analysis that measure language and literacy competencies as well as social and behavioral skills of young children. Instruments are listed with established abbreviations due to limited space.

(Table 14. continued)

Social skills	Assessment, observations, video analysis and teacher and parent ratings involving social-emotional competencies and aspects of positive behavior (ACES accuracy, anger bias, Cooper Farran BRS work-related skills, Creative Curriculum Development Continuum social-emotional development; DPI total score; HS competence scale total score; PBI compliance, social competence; PBS compliance with teacher directives, social competence; PKBS social skills, cooperation, independence, social interactions social withdrawal, PLBS attitude, motivation, persistence, strategy, total score, SCP academic behavior, cooperation, learning, positive behavior, SDQ prosocial skills, TRSSA cooperative participation, school liking, self-directedness, rating on attentiveness, effortful control, positive engagement, prosocial behavior, social competencies, target child behavior constructive engagement and positive social, rating on social activity level, peer play scale, play with objectives)
Problem behavior	Observations, video ratings and teacher and parent ratings involving negative, challenging and problem behavior (BPI externalizing behavior, internalizing problems, CCOF aggression, distractibility, negative effects, noncompliance, CST adaptive problems, ECBI conduct problems, PKBS aggression, anxiety, attention problems, explosive behavior, externalizing composite, internalizing composite, SCP aggression, challenging behavior, distractibility, negative affect, noncompliance, self-control, SDQ behavior difficulties, SESBI conduct problems, TRSSA school avoidance, rating conduct problems, rating following rules and expectations, task behavior, conduct problems, attention problems, video analysis of aggressive behavior, % of observed disruptive behavior, target child behavior on disruptive/aggressive behavior and on negative social behavior)

In addition to the quality rating scales, various instruments measuring child outcomes were categorized. For the meta-analysis on social and behavioral development, 2 categories were made. On the one hand, all outcomes that indicated positive social behavior were aggregated and, on the other hand, all outcomes that measured challenging behavior were synthesized. For the meta-analysis, their direction was recoded to indicate the reduction of negative behavior. The language and literacy outcomes of in-service training programs were categorized by an extensive procedure in which all categories were constructed by 2 dependent coders through discussion. For this purpose, a speech pathologist/therapist that was familiar with the instruments and aspects of language and literacy development acted as an additional coder for this task. First, outcomes were distinguished whether they measured categories of language or literacy development. Second, within these categories, separate subcategories were constructed (see table 14).

3.4. Data extraction and computing effect size

In this investigation, the focus was on contrasts between experimental and control groups as a foundation for effect size estimations. “Contrasts are always associated with only a single degree of freedom (df)” (Hall et al., 1994, p. 22). Within this set of meta-analyses, effect sizes were estimated based on means; studies with independent groups and studies that used

matched data or pre-post-scores were differentiated. From a statistical perspective, there is no technical barrier to estimating effect sizes nor to making interpretations (Borenstein et al., 2009). However, there are several reservations about combining studies with different designs ranging from RCTs that are protected from internal bias to quasi-experimental studies, where effect estimations might be affected by confounding variables. Nevertheless, it is suggested that the quality of studies matters more than their design. In particular, distinctions should be made on “the extent to which the studies are able to yield an unbiased estimate of the effect sizes” (Borenstein et al., 2009, p. 360). Further, meta-analysts can decide if the questions about the relationship between treatment and outcome should be answered by controlling for other factors in RCTs and CRT or not controlling for other factors in observational and quasi-experimental studies. Considering other challenges in addition to comparability and combining studies, 2 features distinguish meta-analysis from other investigations. Raw data from each study is rarely available and meta-analysts depend on statistical summaries reported in papers (Bryk & Raudenbush, 2002). Further, different studies use different outcome measures and raw mean difference scores are only an option when all studies use the same scale (Borenstein et al., 2009; Bryk & Raudenbush, 2002). If different instruments are used, standardized mean differences such as Cohen’s d or Hedges’ g are valuable tools to combine and compare results. Then mean differences are divided by the standard deviation to create an index of standardized mean differences. Standardized mean differences can be seen as mean differences when all data are linearly transformed to a scale, where the standard deviation within groups is equal to one (Deeks, Higgins, & Altman, 2008).

From a statistical point of view, effect sizes can be estimated when a paper reports sufficient information (Borenstein et al., 2009, p. 362). Practical considerations for combining studies are based on the essential requirement that all studies use the same index of treatment effect. The standard approach for continuous outcome measures is to use standardized mean differences as a shared, common metric. Simple transformations are needed to calculate a common metric, but effect estimation has certain requirements on the underlying nature of the data. In addition to the common metric and estimation of the same treatment effect, design-specific estimates of sample variance are needed to reflect the precision of effect sizes (Morris & DeShon, 2002, p. 105).

In particular, combining designs with repeated measures and independent-group designs within a meta-analysis has been discussed extensively. While researchers carrying out primary investigations choose different designs to estimate training effectiveness and there is a pool of relevant intervention studies using within-group and between-group designs (Shadish, Cook, & Campbell, 2002), meta-analysts are faced with concerns about whether results are comparable or not. The literature on meta-analysis does not provide consistent guidance on whether the 2 designs provide equal estimates. Some suggest that repeated measures

designs should be excluded (Lipsey & Wilson, 1993), whereas other meta-analysts combine designs with limited discussion (Gibbons, Hedeker, & Davids, 1993). Certainly, Morris and DeShon (2002) suggest that combining effect sizes across different research approaches “is particularly important when the primary research literature consists of a mixture of independent groups and repeated measures designs” (p. 105). However, combining effect sizes from different designs results in greater added complexity, but there is no statistical reason why changes from baseline outcomes should not be combined with final measure outcomes (Deeks, Higgins, & Altman, 2008; Reeves et al., 2008). To investigate the potential source of bias (e.g., selection, time and differential time effects) from combining designs, a statistical examination can be made through moderator analysis examining experimental designs (Morris & DeShon, 2002).

As mentioned earlier, standardized mean differences are used as a common metric when studies all assess the same outcome, but measure it in different ways (Borenstein et al., 2009, Deeks, Higgins, & Altman, 2008). The formula to calculate standardized mean differences is different for independent and matched groups (see table 15). Cohen’s d is one of the most common tests to calculate effect sizes. The Cochrane Collaboration suggests the use of Hedges’ g for the aggregation of effect sizes (Deeks, Higgins, & Altman, 2008, p. 257). Hedges’ g corrects for bias resulting from small samples (Hedges & Olkin, 1985). To avoid potential sources of bias in treatment effect calculations (Morris & DeShon, 2002), the Comprehensive Meta-Analysis Software Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005) was used to estimate effect size, standard error and variance of treatment effect. In the calculation of standardized mean differences, the effect direction was coded and signs were changed for inherently negative variables (e.g., problem behavior). A common practical problem in including repeated measures is that correlations from pre- to posttest are not reported. Bonate (2000) points out that the average correlation between pretest and posttest measures in psychological research is $r = .60$. Other investigations suggest that there is performance stability over time for the job performance of adults with a pretest-posttest correlation above $r = .50$ (Fleishman & Hempel, 1955; Rambo, Chomiak, & Price, 1983) and for student achievement for average performers with $r = .65$ and for low-performing students with $r = .60$ (Cole, Haimson, Perez-Johnson, & May, 2011). Therefore, a conservative cut-off of $r = .50$ was used to estimate the effect size of change scores.

To compute Hedges’ g and standard error, a two-step procedure was used. First, Cohen’s d , standard error and variance were calculated from a mixture of statistical information and study design. Different formulas are needed for independent and paired groups to calculate effect sizes. Table 15 describes the procedures that were used to calculate Cohen’s d and its standard error, controlling for sufficient data and design as suggested by Borenstein (2009).

Table 15. Formulas for effect size transformation

Dependence	Given information	Formula Standard difference (Cohen's <i>d</i>)	Formula Standard Error of standard difference (Cohen's <i>d</i>)
Independent group	M	RawDiff = Mean1 - Mean2	RawDiffSE = Sqr(1 / N1 + 1 / N2) * SDP
	SD N (per group)	SDP = Sqr((((N1 - 1) * SD1 ^ 2 + (N2 - 1) * SD2 ^ 2) / (N1 + N2 - 2))) StdDiff = RawDiff / SDP	StdDiffSE = Sqr(1 / N1 + 1 / N2 + StdDiff ^ 2 / (2 * (N1 + N2)))
Independent group	N1 N2	StdDiff = RawDiff / SDP	StdDiffSE = Sqr(1 / N1 + 1 / N2 + StdDiff ^ 2 / (2 * (N1 + N2)))
Independent group	N1 N2 t-value	HarmonicN = (2 * N1 * N2) / (N1 + N2) StdDiff = t / (Sqr(HarmonicN) / Sqr(2))	StdDiffSE = Sqr(1 / N1 + 1 / N2 + StdDiff ^ 2 / (2 * (N1 + N2)))
Paired group	Paired means SDPre SDPost Pre-post-correlation NTot	PairedDiff = (Mean2 - Mean1) StdDiff = PairedDiff / (PairedDiffSD / Sqr(2 * (1 - Corr)))	PairedDiffSD = Sqr(SD1 ^ 2 + SD2 ^ 2 - 2 * Corr * SD1 * SD2) PairedDiffSE = PairedDiffSD / Sqr(n) StdDiffSE = Sqr(1 / n + StdDiff ^ 2 / (2 * n)) * Sqr(2 * (1 - Corr))
Paired group	MeanDiff SDDiff NTot Pre-post-correlation	StdDiff = PairedDiff / (PairedDiffSD / Sqr(2 * (1 - Corr)))	PairedDiffSE = PairedDiffSD / Sqr(n) StdDiffSE = Sqr(1 / n + StdDiff ^ 2 / (2 * n)) * Sqr(2 * (1 - Corr))
Paired group ¹⁴	StdDiff (paired) SE NTot	StdDiff = StdPairedDiff * Sqr(2 * (1 - ImputedR))	StdDiffSE = StdPairedDiffSE * Sqr(2 * (1 - ImputedR))

Note. Dependence=statistical information on independent or paired groups; Mean1=mean group one; Mean2=mean group 2; SDP=standard deviation paired; N1=number of group one; N2=number of group 2; SD1=standard deviation group one; SD2=standard deviation group 2; RawDiff=raw difference in means; StdDiff=standardized mean difference, SDP=standard deviation pooled; HarmonicN=harmonic number; NTot=number of total sample; Corr=pre-post-correlation; ImputedR=imputed pre-post-correlation; StdDiffSE=standard error of standardized mean difference; PairedDiffSE=standard error of paired mean difference; RawDiffSE=standard error of raw difference in means.

¹⁴ Cohen's *d* (pooled) is standardized by the pooled SD within groups. Cohen's *d* (change) is standardized by the SD of the differences. Cohen's *d* (changed) was computed and then the value of *d* (pooled) was imputed, assuming a correlation (Imputed *r*) between scores in A and B. Imputed *r* is given as .50.

In a second step, the standardized mean difference for Hedges' g was calculated, by multiplying d with the correction factor J (formula: $g = d * J$). The standard error for Hedges' g was also corrected for J (formula: $SE(g) = SE(d) * J$). The correcting factor J was estimated with the following formula:

$$J = 1 - (3 / (4 * df - 1))$$

Separate formulas were used to estimate correction factors for paired groups ($df = N_{Tot} - 1$) and for independent groups ($df = N_{Tot} - 2$). The correcting factor J is always smaller than 1.0 and therefore Hedges' g will always be less than Cohen's d , but J will be very close to 1.0 unless df is very small (Borenstein et al., 2009, p. 28).

3.5. Aggregation of findings and moderator analysis

Two considerations underlie the aggregation of findings and moderator analysis. First, a random effect model incorporates the assumption that different studies estimate different intervention effects and between-study variance emerges. In particular, confidence intervals and aggregated effect sizes are wider in a heterogeneous data set. Estimates of statistical significance are more conservative if random effect models are used (Deeks, Higgins, & Altman, 2008). Within a random effect model, it is assumed that the effect in different studies is similar, but not identical, and follows the same distribution. In particular, this occurs when studies differ in the mixture of participants and in intervention implementation, use different measures and for other reasons (Borenstein et al., 2009). Assessing the impact of educational interventions, the magnitude of effects might vary depending on characteristics of the educational settings such as class size, age of children, or proportions of disadvantaged children in the classroom. Educational setting factors are likely to vary from study to study. Therefore, random effect models were chosen for the theoretical reason that different population distributions can be assumed due to the fact that the studies included were based on various outcome measures and multiple in-service training approaches in various early childhood settings. Second, the use of a multi-level approach acknowledges the hierarchical structure of effect sizes nested under studies (Bryk & Raudenbush, 2002). For example, in measuring the impact of in-service programs on language abilities, various specific categories of language were assessed within studies (receptive or expressive vocabulary, comprehension). To aggregate such different outcome measures into a summary effect size, a multi-level approach is suggested (Bryk & Raudenbush, 2002; Hedges, 2009). Specifically, a 2-level approach was used to consider variations in effect size and treatment level.

To calculate the non-weighted meta-analysis and moderator analysis, MLwiN (Rasbash, Charlton, Browne, Healy & Cameron, 2005) was used. Effect size aggregation and tests for

heterogeneity and sensitivity were conducted with the Comprehensive Meta-Analysis Version 2 (Borenstein et al., 2005) using a random effect model with an aggregated weighted effect size (Hedges' g'). To calculate a precise summary effect size with minimum variance, weighted means were calculated, where "the weight assigned to each study is the inverse of the study's variance" (Borenstein et al., 2009, p. 72). In random effect models, there are 2 sources of variance. First, there is within-study error in estimating the effect in each study, because the observed value Y for any study i differs from the study's true value (θ_i). Second, there is some variation in the true effect size, because the true value θ_i for each study differs from μ due to between-study variance τ^2 . The between-study variance τ^2 is the sum of the total variance Q minus degrees of freedom ($df = k - 1$) divided by the denominator C . The denominator C is a scaling factor that ensures that τ^2 is in the same metric as the within-study variance.

$$\tau^2 = \frac{Q-df}{C} \quad Q = \sum_{i=1}^k W_i Y_i^2 - \frac{(\sum_{i=1}^k W_i Y_i)^2}{\sum_{i=1}^k W_i} \quad C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

The study variance for the random effect model V'_{yi} is the within-study variance V_{yi} for study i plus the between-study variance τ^2 . Weights (W'_i) were assigned under the random effects model to each study as followed:

$$W'_i = \frac{1}{V'_{yi}}$$

The weighted summary effect size g' is the sum of effect size Y_i multiplied by the weight W'_i and divided by the sum of the weights W'_i .

$$g' = \frac{\sum_{i=1}^k W'_i Y_i}{\sum_{i=1}^k W'_i}$$

The variance of the summary effect is reciprocal to the sum of the weights and the standard error of the summary effect is the square root of the variance.

$$Vg' = \frac{1}{\sum_{i=1}^k W'_i} \quad SEg' = \sqrt{Vg'}$$

The lower and upper 95 % confidence interval for the summary effect sizes was computed as follows:

$$LLg' = g' - 1.96 * SEg' \quad ULg' = g' + 1.96 * SEg'$$

Heterogeneity and stability of aggregated effect size

For all sets of data in the meta-analysis, a number of statistics on heterogeneity and sensitivity were estimated to justify the use of a random model. The Q-statistic was used as test of homogeneity. Q is sensitive to the ratio of observed variation in within-study error. "The Q

test is computed by summing the squared deviations of each study's effect estimate from the summary effect estimate, weighting the contribution of each study by its inverse variance" (Huedo-Medina et al., 2006, p. 4). When all studies share a common effect size (null hypothesis), Q follows a chi-square distribution with the degree of freedom of $k-1$, where k is the number of studies (Deeks, Higgins, & Altman, 2008). A significant p-value (rejecting the null hypothesis) provides evidence for some variance in the true effect sizes and supports the rationale for applying a random-effect model (Huedo-Medina et al., 2006). The I^2 -statistic represents the proportion of observed variance that reflects real differences in effect sizes [Formula: $I^2 = ((Q^2 - df) / Q) \times 100$]. It reflects the extent of overlap of confidence intervals and quantifies the inconsistencies¹⁵ (Deeks, Higgins, & Altman, 2008). Further, between-study variance in effect sizes was estimated with Kendall's τ^2 (cf. Borenstein et al., 2009).

Sensitivity analysis was conducted to estimate the stability of the average effect size. In particular, a one-study-remove analysis was used to estimate the stability of the average effect size, where one study was removed to identify outlier studies. In addition, Classical and Orwin Fail Safe N analysis was carried out to calculate how many studies/treatments with null results (mean effect of zero) are needed to bring the average effect down to a nonsignificant value of $g < 0.20$ or 0.

Publication bias

Meta-analysis aims at a mathematically accurate synthesis of a set of studies. If the studies are based on a biased selection of investigations then the computed effect will reflect this bias. The issue that published studies are more likely to be integrated in meta-analyses is called publication bias or the file drawer effect (cf. Borenstein et al., 2009; Rosenthal, 1979). In addition, small studies without significant effects are more likely to remain unpublished. Further, the tendency for effect size estimates in small studies to differ from those in large studies is called the small-study effect. In particular, poor methodological quality leads to inflated effects in smaller studies (cf. Sterne, Egger, & Moher, 2008). The precision of effect size estimates increases as the size of studies increases (cf. Harbord, Harris, & Sterne, 2009). Graphs, for example, or box and funnel plots, can be used to display the relationship between sample size or standard error (y axis) and effect size (x axis). A file drawer problem is assumed when there is asymmetry with many small studies lying at the bottom of the graph and spread around the full range of the graph (cf. Borenstein et al., 2009; Harbord, Harris, & Sterne, 2009). Because interpreting plots seems to be largely subjective, Egger regression is helpful to investigate the file drawer problem through a linear regression approach to measure funnel plot asymmetry (Borenstein et al., 2009; Egger, Smith, Schneider,

¹⁵ Interpretation of I^2 -statistic according to Deeks, Higgins, & Altman, 2008, p. 278: 0-40% might not be important; 30-60% represent moderate heterogeneity; 50-90% represent substantial heterogeneity; 75-100% represent considerable heterogeneity.

& Mind, 1997c). It is proposed that researchers “perform a linear regression of the intervention effect estimates on their standard error, weighted by $1 / (\text{variance of the intervention effect estimates})$ ” (Sterne, Egger, & Moher, 2008, p. 314). A straight-line relation between intervention effects and standard errors is assumed and the intercept provides a measure of asymmetry, where a larger deviation than zero indicates a more pronounced asymmetry (Egger et al., 1997c).

Moderator analysis

The variation in outcomes across studies can be analyzed through meta-regression which tests for potential moderators (Bryk & Raudenbush, 2002). In addition to the aggregation of research results, meta-regression is used to examine theoretical moderators and potential methodological effect modifiers. Meta-regression, an extension of subgroup analysis, allows researchers to investigate the effect of a single moderator or multiple factors simultaneously. The outcome variable (effect size) is predicted by one or more explanatory variables. In particular, potential effect modifiers (characteristics of studies) that predict treatment effects can be examined, when there are more than 10 studies included (Deeks, Higgins, & Altman, 2008). The obtained regression coefficient describes how the intervention effect changes with a unit increase in the explanatory variable (effect modifier) and the statistical significance shows the linear relation between effect and modifier. Dummy coded categories can also be used to investigate differences for subgroups. Then the regression coefficient shows how the intervention effect differs between reference groups and the significance test indicates statistically considerable differences (Deeks, Higgins, & Altman, 2008, p. 285). The Cochrane Collaboration suggests that heterogeneity among studies should be investigated through meta-regression at least for baseline risk characteristics and dose-response aspects (ibid, pp. 288-289). Restricted maximum likelihood (RML) estimations are suggested by Hox (2010) to examine the between-study variance. Restricted maximum likelihood (RML) estimations are used, because a) variance is viewed as a nuisance parameter in multilevel approaches and it is important to include it in the model and b) they provide a good estimate of between-study variance and significance. The aggregation of effect sizes and the moderator-analysis (meta-regression) were carried out with MLwiN (Rasbash et al., 2005). A two-level analysis, described by Hox (2010), was carried out to include treatment number, effect size (Hedges' g), standard error of effect size and regression constant. Outcomes were categorized at first level and treatments at second. The standard error of effect sizes (predictor) was included in the random part at level one, with a fixed coefficient at value 1. The regression constant was involved in random and fixed parts at level 2. In the moderator analysis, explanatory variables were included in the fixed part only (Hox, 2010, p. 231). Further, a chi-square test is suggested to test residual variance (cf. Bryk & Raudenbush, 2002; Hox, 2010).

4. Results

4.1. Results on the impact on quality ratings

This section deals with the research question to what extent external quality ratings in early childhood settings can be improved through in-service professional development. It describes differences in the magnitude of effects when different quality constructs and scales are used (i.e., CLASS, ERS, ELLCO). Through multi-level meta-regression, training duration and intensity as potential modifiers are investigated. Additionally, it is analyzed whether specific professional development delivery formats are more effective than others. The literature hypothesizes that coaching or scale-based in-service approaches are more effective than others. Further, it is supposed that professional development is less effective when it is brought to scale due to the heterogeneity of trainers and conditions. Additionally, the workplace environment and curriculum are assumed to be potential effect modifiers.

4.1.1. Description of primary studies on quality ratings

Overall 36 studies with 42 different in-service treatments reported appropriate statistical data to conduct a meta-analysis. Approximately one third of studies ($n = 12$) were unpublished (e.g., research reports, conference papers and dissertations) and 2 thirds were published studies ($n = 24$). Almost all the studies were conducted in the United States, except 2 from Canada (Howe et al., 2011; Japel, 2009). Most of the studies ($n = 19$) were published recently, within the last 5 years. The majority of studies ($n = 19$) had experimental designs: RCT with randomization at the individual level ($n = 11$) and cluster randomized trials ($n = 8$) with randomized assignment of classrooms or centers. Further, 9 studies applied a quasi-experimental design with non-randomized assignment of participants, and 8 investigations used a non-experimental design evaluating the progress of a treatment group without a control condition. Overall, data from 2,891 teachers were included in the analysis. The sample sizes varied from 6 to 553 educators at the beginning of the studies. To systemize and compare the professional development approaches, the delivery formats of the training programs were coded into the following categories: workshop, coursework, online and distance training, community of practice, and onsite support. As can be seen in table 16, approximately 2 thirds ($n = 23$) of in-service approaches applied multiple delivery formats. One third of in-service programs used a single delivery strategy, either a workshop ($n = 4$), a course ($n = 4$) or onsite support ($n = 5$). The duration of the professional development approaches

ranged from very short programs with half-day workshops or 4 days of coaching (e.g., Burchinal et al. 2002; Englund, 2010; Japel, 2009) to 3 years' long programs. Most professional development programs focused on language and literacy (i.e., 14 studies with 17 different treatments). In all, 8 studies with 9 treatments focused on quality improvements in early childhood classrooms. In 9 studies and 11 treatments, curriculum implementation was the major focus. The other studies emphasized general topics of early childhood education ($n = 2$), leadership, the social-emotional development of young children or science education. Most of the studies ($n = 35$) utilized internationally renowned quality rating scales. In summary, 17 studies measured the impact of professional development programs at classroom level with the ELLCO, 12 with the ERS, and 9 with the CLASS. Almost all the programs that focused on language and literacy themes measured the impact with the ELLCO (16 of 17) and nearly all quality improvement approaches applied the ERS (8 of 9). Overall, training amounts ranged from 4 hours to 308 hours.

Eight studies (reporting 9 treatments) evaluated scale-based training sessions. Within 13 studies with 16 treatments, a curriculum or a set of specific manual-based activities/sessions were included in the professional development. However, only 9 studies with 11 treatments reported that the professional development focused on curriculum implementation as the major topic. Overall, 289 effect sizes could be extracted from contrasts between experimental and control groups at posttest or change scores between pretest and posttest on external quality ratings. Effect sizes (see figure 6) ranged from a strongly negative impact ($g = -0.82$) to giant positive effects ($g = 6.62$). Approximately one quarter of the effect sizes ($k = 73$) demonstrated no or negative impacts ($g < 0.20$), 68 were small effects ($g > 0.20$), 62 were medium effect sizes ($g > 0.50$) and approximately 30 % of effect sizes ($k = 86$) revealed large effects ($g > 0.80$) of in-service programs on quality ratings, according to Cohen's rules of thumb (1988).

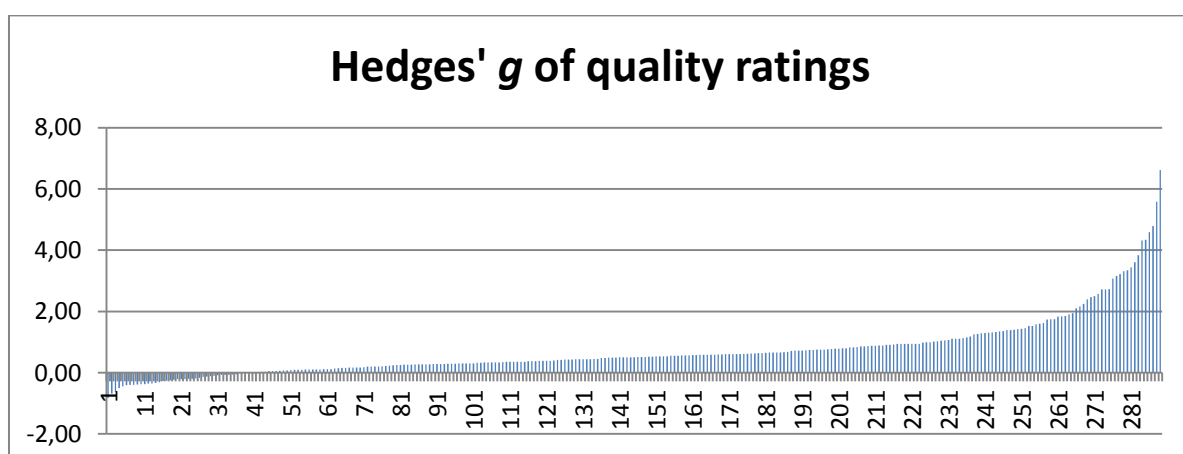


Figure 6. Effect size distribution of quality ratings.

Table 16. Overview of studies included in meta-analysis on quality ratings

Author	Year	Country	Pubtype	Design	Sample teacher	WS	CS	ONL	COP	ON	Months	Topic	Scale	Hours
Algozzine et al.	2011	USA	pub	CRT	72		x			x	12.00	CUR	ELLCO	360
Barnett et al.	2008	USA	pub	RCT	18	x				x	9.00	CUR	ERS,CLASS,SELA	59
Bloom & Sheerer	1992	USA	pub	QNE	44		x				16.00	LEAD	ECCOS	308
Boller et al.	2010	USA	unpub	RCT	129					x	7.00	QI	ERS	56
Breffni	2011	USA	pub	RCT	20		x				2.00	CUR	CLASS	32
Brinks	2007	USA	unpub	OGPP	31	x	x	x		x	9.00	L&L	ELLCO	-
Bryant et al.	2009	USA	unpub	RCT	108					x	9.00	QI	ERS	30
Burchinal et al.	2002	USA	pub	QNE	553	x					0.10	GEN	ERS	-
Buyse, Castro, & Peisner-Feinberg	2010	USA	pub	CRT	55	x			x	x	2.00	L&L	ELLCO	40
Campbell & Milbourne	2005	USA	pub	QNE	180		x			x	3.50	QI	ERS	18
Cassidy et al.	1995	USA	pub	QNE	41		x				7.50	GEN	ERS	-
Clancy-Menchetti	2006	USA	unpub	RCT	77	x	x			x	24.00	L&L	ELLCO	276
Cunningham ¹⁶ (a)	2007	USA	unpub	RCT	304		x				3.00	L&L	ELLCO	30
Cunningham ⁷ (b)	2007	USA	unpub	RCT	304		x			x	9.00	L&L	ELLCO	94
Dickinson & Caswell	2007	USA	pub	QNE	70		x			x	5.00	L&L	ELLCO, AP	45
Domitrovich et al.	2009	USA	pub	CRT	88	x				x	9.00	CUR	CLASS	68
Edgar ¹⁷	2008	USA	unpub	OGPP	94				x	x	2.25	L&L	ELLCO	18
Englund	2010	USA	unpub	OGPP	6					x	0.13	L&L	CLASS	4
Grace et al.	2008	USA	pub	CRT	40		x			x	36.00	L&L	ELLCO	-
Howe et al. (a)	2011	Canada	pub	QNE	94					x	3.75	CUR	ERS	45
Howe et al. (b)	2011	Canada	pub	QNE	94	x					3.75	CUR	ERS	8
Japel (a)	2009	Canada	unpub	CRT	30	x				x	12.00	QI	ERS	-
Japel (b)	2009	Canada	unpub	CRT	30	x					0.10	QI	ERS	-
Lonigan et al. (a)	2011	USA	pub	CRT	51	x				x	9.00	CUR	ELLCO,ECCOM	68
Lonigan et al. (b)	2011	USA	pub	CRT	51	x					9.00	CUR	ELLCO,ECCOM	32
McNerney, Nielsen, Clay	2006	USA	pub	OGPP	23	x				x	6.00	L&L	ELLCO	-
Morris et al.	2010	USA	unpub	CRT	51	x	x			x	9.00	CUR	(in)CLASS	162
Neuman & Cunningham (a)	2009	USA	pub	RCT	304		x				3.75	L&L	ELLCO	45
Neuman & Cunningham (b)	2009	USA	pub	RCT	304		x			x	8.00	L&L	ELLCO	85

¹⁶ Cunningham (2007) and Neuman & Cunningham (2009) use data from the Project Great Start Professional Development Initiative, but report different results.

¹⁷ Study of Edgar (2008) compares three different coaching approaches; however only for master coaching statistical data for meta-analysis is available.

(Table 16. continued)

Author	Year	Country	Pubtype	Design	Sample teacher	WS	CS	ONL	COP	ON	Months	Topic	Scale	Hours
Neuman & Wright (a)	2010	USA	unpub	RCT	148		x				2.50	L&L	ELLCO	30
Neuman & Wright (b)	2010	USA	unpub	RCT	148					x	2.50	L&L	ELLCO	30
Onchwari & Keengwe	2010	USA	pub	QNE	44					x	5.00	L&L	ELLCO	40
Palsha & Wesley	1998	USA	pub	OGPP	73	x				x	7.68	QI	ERS	30
Powell et al. (a)	2010	USA	pub	RCT	88	x				x	9.00	L&L	ELLCO	-
Raver et al.	2008	USA	pub	RCT	94		x			x	7.00	SOC	ERS, CLASS	132
Roehrig et al.	2011	USA	pub	OGPP	37	x	x	x			24.00	SCI	CLASS	-
Sheridan	2006	USA	pub	QNE	61		x			x	12.00	QI	ERS	-
Sibley & Sewell	2011a	USA	pub	QNE	72	x	x		x	x	36.00	CUR	ELLCO	-
Sipp	2010	USA	unpub	OGPP	24		x			x	1.50	QI	CLASS	12
Sprague et al.	2009	USA	unpub	RCT	20		x			x	5.00	CUR	ELLCO,ELA,LOC	105
Uttley & Horm	2008	USA	pub	OGPP	21	x	x			x	8.00	QI	ERS	-
Wasik & Hindman	2011	USA	pub	CRT	30	x	x			x	9.00	L&L	ELLCO,CLASS	50

Note. Pubtype=publication type; pub=published in journals or as book chapter; unpub=unpublished paper e.g., technical paper, dissertation or governmental report; design=study design; WS=workshop; CS=course and in-service with ongoing sessions; ONL=online training or blended learning; COP=community of learning or community of practice; ON=onsite support like coaching, consulting etc.; Topic=focus of training; CUR=curriculum implementation; L&L=language and literacy; SOC=social and behavioral development; QI=quality improvement initiative; SCI=science; GEN=general focus of early childhood education(e.g., developmental appropriate practice); ; RCT=randomized control trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center to groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design; OGPP=one group pre-post measure; AP=Assessment Profile Learning Environment; CLASS=Classroom Assessment Scoring System; ECCOM=Early Childhood Classroom Observation Measure; ECCOS=Early Childhood Classroom Observation Scale; ELA=Early Literacy Activity Rating Scale; ELLCO=Early Language and Literacy Classroom Observation; ERS=Environmental Rating Scales (e.g., ECERS or ITES); (in)CLASS=Individualized Classroom Assessment Scoring System; LOC=Literacy Observation Checklist, SELA=Support for Early Literacy Assessment.

A test for heterogeneity, conducted with the CMA (Borenstein et al., 2005), indicated a between-study variance ($Q = 746.108$; $df = 41$; $p < .000$; $T^2 = .189$; $SE = .072$) with a high percentage (95 %) of observed variance across treatments due to systematic variance ($I^2 = 94.51$). The heterogeneity statistics led to the decision to use a random effects model and to conduct a meta-regression. Orwin's Fail Safe N analysis showed that at least 50 studies with a mean effect size of zero are needed to decrease the summary effect size to a negligible impact of $g < 0.20$. The test of sensitivity demonstrated a robust summary effect that slightly varied between $g = 0.62$ to $g = 0.71$ if one treatment was removed. Considering the file drawer aspect, chi-square analysis indicated that the effect size variance was not significantly related to sample size of teachers ($\chi^2 = 7446$; $df = 7280$; $p = .085$) or publication type ($\chi^2 = 289$; $df = 200$; $p = .343$). A chi-square test of effect size and standard error ($\chi^2 = 75938$; $df = 74480$; $p < .000$) and Egger regression (Intercept = 3.91; $SE = 1.08$; $p < .000$) suggested an asymmetry, where small studies with high effects were missing.

4.1.2. Meta-analysis and moderator analysis on quality ratings

The meta-analysis of 36 studies with 42 treatments estimated a significant medium-weighted summary effect size ($g' = 0.68$; $SE = 0.071$; $k = 289$; $CI = 0.55$ to 0.82 ; $Z = 9.696$; $p < .000$) for in-service programs on quality ratings. The confidence interval for the weighted summary effect was relatively narrow in the medium effect size range. Within the meta-analysis, studies were weighted according to their variance, ranging from 0.55 to 2.79 %. The aggregated effect size per treatment, displayed in table 17, ranged from a small negative impact ($g = -0.39$) to huge positive effects ($g = 3.66$). The confidence intervals for all treatments were relatively wide, indicating that the treatment effects were not very accurate. The lower confidence interval estimates for most of the treatments were positive, except for 9 treatments. The width of the confidence intervals per treatment suggests that there are different degrees of benefit from in-service professional development for teachers.

Conducting an unweighted meta-analysis and multi-level meta-regression with MLwiN, where effect size was nested under treatments, the summary effect was $g = 0.73$ ($SE = 0.122$; $k = 289$; $RIGLS = 566.6$; $\Delta\chi^2(41) = 1616$; $p < .000$). Through multi-level moderator analysis, methodological and theoretical effect modifiers were tested in model 1 and methodological artefacts were included as covariates in model 2 to empirically confirm theoretical moderators.

Table 17. Meta-analysis on quality ratings

Nr.	Study	ES	SE	Lower CI	Upper CI	Relative weight	Hedges' g and 95% CI
1	Algozzine et al.	0.59	0.11	0.37	0.80	2.63	1
2	Barnett et al.	0.63	0.12	0.40	0.86	2.61	2
3	Bloom & Sheerer	1.07	0.14	0.80	1.35	2.52	3
4	Boller et al.	3.66	0.31	3.05	4.28	1.81	4
5	Breffni	3.03	0.35	2.35	3.71	1.67	5
6	Brinks	2.10	0.20	1.71	2.49	2.30	6
7	Bryant et al.	0.21	0.11	-0.01	0.42	2.63	7
8	Burchinal et al.	0.36	0.05	0.26	0.45	2.77	8
9	Buyse et al.	0.17	0.11	-0.04	0.39	2.63	9
10	Campbell & Milbourne	0.17	0.19	-0.19	0.54	2.35	10
11	Cassidy et al.	0.50	0.20	0.11	0.89	2.30	11
12	Clancy-Menchetti	-0.39	0.31	-1.01	0.22	1.81	12
13	Cunningham (a)	0.38	0.18	0.02	0.74	2.37	13
14	Cunningham (b)	0.86	0.19	0.49	1.23	2.34	14
15	Dickinson & Caswell	1.04	0.08	0.88	1.20	2.71	15
16	Domitrovich et al.	0.40	0.10	0.22	0.59	2.67	16
17	Edgar	0.39	0.05	0.29	0.48	2.77	17
18	Englund	2.72	0.86	1.04	4.40	0.55	18
19	Grace et al.	2.42	0.17	2.09	2.75	2.42	19
20	Howe et al. (a)	1.10	0.28	0.56	1.64	1.97	20
21	Howe et al. (b)	-0.06	0.26	-0.56	0.44	2.05	21
22	Japel (a)	0.49	0.16	0.17	0.82	2.44	22
23	Japel (b)	0.40	0.16	0.09	0.72	2.45	23
24	Lonigan et al. (a)	0.32	0.10	0.12	0.52	2.65	24
25	Lonigan et al. (b)	0.21	0.10	0.00	0.41	2.65	25
26	McNerney et al.	0.18	0.07	0.04	0.33	2.73	26
27	Morris et al.	0.18	0.04	0.11	0.25	2.79	27
28	Neuman & Cunningham (a)	0.31	0.07	0.18	0.45	2.74	28
29	Neuman & Cunningham (b)	0.68	0.07	0.54	0.82	2.73	29
30	Neuman & Wright (a)	0.15	0.09	-0.02	0.33	2.69	30
31	Neuman & Wright (b)	0.54	0.09	0.36	0.72	2.68	31
32	Onchwari & Keengwe	0.45	0.17	0.11	0.80	2.40	32
33	Palsha & Wesley	0.86	0.09	0.69	1.04	2.69	33
34	Powell et al. (a)	1.39	0.19	1.03	1.75	2.36	34
35	Raver et al.	0.64	0.11	0.43	0.84	2.64	35
36	Roehrig et al.	0.64	0.10	0.43	0.84	2.65	36
37	Sheridan	1.29	0.49	0.34	2.24	1.20	37
38	Sibley & Sewell	0.31	0.12	0.08	0.55	2.60	38
39	Sipp	-0.22	0.08	-0.37	-0.07	2.72	39
40	Sprague et al.	0.34	0.18	-0.01	0.68	2.39	40
41	Uttley & Horm	0.48	0.25	-0.01	0.97	2.08	41
42	Wasik & Hindman	1.71	0.31	1.11	2.31	1.84	42
43	Summary effect	0.68	0.07	0.55	0.82		43

First, the methodological variables were examined through a multi-level moderator analysis and subgroup analysis. In table 18, the direct effect size estimates of the subgroup analysis for dichotomous moderators are displayed. For continuous interval scaled moderators, the regression coefficients that indicate a change in effect size with a unit increase of the exploratory moderator variable are reported. In model 1, all exploratory moderator variables are analyzed individually. Looking at publication type, it became apparent that the effect size of unpublished studies was considerably smaller than for published ones. Studies published before and after the millennium reported similar effect sizes. As regards methodological issues, the use of different quality rating scales matters. In particular, impacts were significantly higher when renowned scales ($g = 0.76$), like CLASS, ERS, or ELLCO, were applied instead of newly developed/researcher developed rating instruments ($g = 0.41$). Comparing the renowned scales, outcomes using the CLASS had higher values ($g = 1.04$), while outcomes measured with the ERS or ELLCO were in line with the summary effect. All reported outcomes were coded for instrument reliability and differences in mean value at pretest at effect size level. Meta-regression demonstrated that the outcomes, where the experimental groups had slightly higher values than the control groups at the beginning resulted in higher effect sizes at posttest ($g = 0.99$). Further, effect sizes were significantly smaller for outcomes ($g = 0.53$), where teachers in the control group had lower values in quality ratings at pretest. The effect size ($g = 0.99$) of outcomes with a highly reliable quality rating scale (Cronbach's $\alpha > .70$) was significant higher in comparison to outcomes, where lower reliability scores or no reliability values were given. Inter-rater reliability was also coded as a potential effect modifier and effect size was smaller for unreliable instruments. Effects were not statistically significantly modified by treatment comparison type or the assignment mechanism of teachers to treatment or control groups. Regarding trainee characteristics, in-service training effectiveness does not differ when the sample consists exclusively of teachers with a university qualification or mixed qualification levels.

The background variables of child care centers and classrooms were investigated to see if they moderate training effectiveness. Information on the number of children at risk in classrooms and children under the age of 3 was successfully coded. However, neither children at risk nor children under the age of 3 influenced the effect size significantly.

Training intensity and in-service duration are hypothesized as being effect modifiers in the literature. However, meta-regression did not support the suggestion that a longer duration (*regression coefficient* = .010; SE = 0.015) and higher in-service training amounts in hours (*regression coefficient* = -.001; SE = 0.002) led to higher in-service training effects in general. Splitting duration and amount, the results showed that in-service programs ($n = 6$) with a training amount of 45 to 60 hours ($g = 1.93$) were significantly more effective than other programs.

Investigating the *in-service delivery format*, in-service programs that used coaching as the sole training strategies ($g = 1.98$) were almost 3 times more effective than other programs ($g = 0.67$). The findings also suggested that providing onsite support only or in combination with other delivery formats (e.g., workshops or courses) does not make training automatically more effective. These onsite models might not be systematic and manual-driven enough to moderate improvements in child care quality sustainably. Further, using multiple delivery formats was not found to be necessarily more successful than formats using a single training strategy.

In the literature, it has been hypothesized that *scale-based professional development* approaches guide in-service training providers on their work with teachers and provide feedback to teachers on material purchasing and instructional sequencing and delivery (McNerney et al., 2006). A moderator analysis (*regression coefficient* = .435; SE = 0.295; $k = 279$) did not support the assumption that scale-based training is notably more effective than training that is not based on quality ratings or performance scales. The *curriculum* as a potential moderator was suggested by Klein and Gomby (2008). Moderator analysis did not provide significant evidence that the implementation of a comprehensive or specific curriculum through professional development led to statistical considerable changes in early childhood education quality. Moderation was tested whether curriculum implementation was major target of the program (*regression coefficient* = -.169; SE = 0.278, $k = 289$) or curriculum/manual based strategies were part of the training in combination with other training aspects (*regression coefficient* = -.172; SE = 0.260; $k = 289$).

So far, the meta-regression did not provide evidence that large scale professional development is less effective than small scale programs, using the number of providers (*regression coefficient* = -.004; SE = 0.007, $k = 120$) and number of training participants (*regression coefficient* = -.001; SE = 0.001; $k = 289$) as large scale indicators. The training effect did not decrease systematically when the numbers of participants or trainers increased. However, information on the number of in-service providers was only available for some studies.

Model 1 showed that the effect size of the summary effect is significantly influenced by several methodological artefacts. In particular, the effect size is larger a) when renowned scales were used (1=renowned scales [CLASS; ERS; ELLCO]; 0=not) and b) when high reliability values of single outcome were reported (1= $\alpha > .70$) and significant smaller c) when the control group had higher values at pretest. These methodological artefacts were used as covariates in model 2 to tested whether the theoretical effect modifiers remained as significant predictors or not, adjusting for methodological artefacts. The 3 covariates were entered in the equation simultaneously and all 3 remained as significant predictors. The correlations between the 3 covariates of model 2 were modest ($r < .21$).

In model 2, a theoretical effect modifier was entered simultaneously in the equation in addition to the 3 covariates. After controlling for covariates, the moderator analysis confirmed the findings that delivery format and in-service program intensity matter. In particular, the findings on professional development programs with an intensity of 45 to 60 hours and programs that provided coaching solely were validated in model 2. The findings on publication type could not be confirmed, adjusting for methodological characteristics.

Table 18. Moderator analysis of in-service training effects on quality ratings

Characteristic	Model 1				Model 2	
	<i>k</i>	<i>g</i>	<i>k</i> =289 Residual variance	$\Delta\chi^2$	<i>k</i> =289 Significance of moderator	Residual variance
Publication type						
Published	195	0.75	.597	780*	n.s.	.579
Unpublished	94	0.69*	.597	780*	n.s.	.579
Published after 1999	261	0.75	.598	1447*	n.s.	.585
Methodology						
Renowned scale						
Yes	260	0.76*	.579	1303*	-	-
No	29	0.41*	.579	1303*	-	-
CLASS	66	1.04*	.580	248*	n.s.	.563
ELLCO	127	0.77	.590	740*	n.s.	.573
ERS	67	0.85	.594	279*	n.s.	.561
Difference at pretest						
Experimental group higher	101	0.85*	.565	447*	-	-
Control group higher	25	0.53*	.567	156*	-	-
Reliability						
Cronbach's $\alpha > .70$	92	0.99*	.589	470*	-	-
Below or not reported	197	0.64*	.589	470*	-	-
Inter-rater reliability						
<i>Kappa</i> or <i>ICC</i> $\geq .80$	157	0.81	.693	903*	n.s.	.588
<i>Kappa</i> or <i>ICC</i> below/not reported	132	0.68*	.603	903*	n.s.	.588
Treatment comparison type						
Contrast EG vs. CG	207	0.73	.601	1075*	n.s.	.585
Change from pre to post	82	0.75	.601	485*	n.s.	.585
Random assignment	181	0.74	.599	948*	n.s.	.585
Teachers with university degree	46	0.63	.599	243*	n.s.	.584
Classroom composition/setting						
At risk population (children)	204	0.65	.596	1021*	n.s.	.585
Children under age 3	57	0.71	.600	319*	n.s.	.584
In-service training intensity						
Amount						
Less than 30 hours	53	0.44	.562	410*	n.s.	.544
30 to < 45 hours	43	0.85	.597	210*	n.s.	.580
45 to < 60 hours	28	1.93*	.484	139*	sig.	.489
60 to < 100 hours	30	0.45	.591	150*	n.s.	.572
100 to < 120 hours	6	0.34	.596	5	n.s.	.580
120 to < 170 hours	27	0.41	.594	212*	n.s.	.582
> 170 hours	16	0.46	.596	34	n.s.	.584
Duration						
Less than 3 months	50	0.69	.598	499*	n.s.	.580
3 to < 6 months	61	0.44	.580	241*	n.s.	.563
6 to < 9 months	129	0.87	.584	662*	n.s.	.572
9 to < 12 months	19	0.74	.599	24	n.s.	.583
12 to < 24 months	19	0.47	.596	26	n.s.	.584
24 to 36 months	11	1.35	.574	108*	n.s.	.555

(Table 18 continued)

Characteristic	Model 1 <i>k</i> =289				Model 2 <i>k</i> =289	
	<i>k</i>	<i>g</i>	Residual variance	$\Delta\chi^2$	Significance of moderator	Residual variance
In-service training delivery						
Onsite	226	0.76	.597	1281*	n.s.	.583
Coaching	79	0.11	.541	465*	n.s.	.553
Consultation	55	0.37	.577	313*	n.s.	.573
Mentoring	45	0.52	.595	172*	n.s.	.581
Only coaching						
Yes	14	1.98*	.520	123*	sig.	.524
No	275	0.67*	.520	123*	sig.	.524
Onsite support only						
Yes	41	1.20	.568	190*	n.s.	.559
No	248	0.64	.568	190*	n.s.	.559
In-service training delivery						
Scale-based training	94	0.62	.593	420*	n.s.	.577
Multiple delivery formats	186	0.65	.595	1091*	n.s.	.575
Curriculum implementation						
Mayor focus of training	102	0.61	.594	384*	n.s.	.582
Content part of the training	118	0.72	.596	434*	n.s.	.581

Note. * $p < .05$; n.s.=not significant; sig.=significant; model 2. Significance tested with methodological covariates (renowned instrument, high reliability, control group score lower at pretest).

To sum up, medium-sized to large summary effects ($g' = 0.68$; $g = 0.73$) were found for the impact of professional development programs on external quality ratings. The moderator analysis suggested that the impact was affected by the quality scale or construct that was used. In particular, results that were measured with renowned scales were larger. Comparing the 3 renowned scales, the only outcomes that were evaluated with the CLASS were rated significantly larger. There was no significant difference between outcomes measured with the ERS or ELLCO in comparison to the summary effect size. Analyzing baseline risk factors, only the pretest values of the control or experimental groups were found to be a significant moderator. Previous pre-service education and qualification levels of teachers did not appear as a significant effect modifier. Relationships between in-service effects and amount of training could be confirmed. Meta-regression found a relationship between in-service training effects and a specific amount (45 to 60 h). No linear relationship between amount of training and in-service training effects appeared that would have indicated an automatically effect size increase per training hour. Further, coaching as sole delivery format was found to be the most promising approach to improve child care quality sustainably. So far, there is no pivotal evidence that training effectiveness is improved when scale-based approaches are used or when in-service programs are brought to scale. Further, curriculum implementation does not seem to influence training impacts on quality ratings. The workplace environment, coded as classroom composition including age group and at risk population, was not found to moderate in-service impacts statistically significantly.

4.1.3. Narrative review of effective studies on quality ratings

Overall, 13 in-service treatments were found to have a large aggregated effect on quality ratings ($g > 0.80$). Comparing the 13 most effective in-service programs by structural features, there was great variety in sample size, training amount and duration as well as delivery format. Sample sizes ranged from 6 to 304 teachers, with a mean of 76 teachers. Further, short-term programs with a 4-day onsite procedure (Englund, 2010) were equally as effective as long-term approaches with a duration of up to 3 years (Grace et al., 2008). Great variability was also found for training amounts, ranging from 4 to 308 training hours, with a mean of 73 hours. Overall, 2 treatments used coursework (Bloom & Sheerer, 1992; Breffni, 2011), 3 onsite support only (Boller et al., 2010; Englund, 2010; Howe et al., 2011), 3 combined workshop and onsite support (Palsha & Wesley, 1998; Powell et al., 2010; Sheridan, 2006), 3 used course and onsite support (Dickinson & Caswell, 2007; Cunningham, 2007; Grace et al., 2008), and the others combined 3 or more delivery formats (Brinks, 2007; Wasik & Hindman, 2011).

Regarding *financial support*, the largest aggregated impact ($g = 3.76$) was found for the Seeds to Success Program by Boller et al. (2010). The program comprised of 8 hours of coaching per month per classroom, quality improvement grants of up to 12,000 USD, and funds for professional opportunities and additional support for child care expenses, release time and books. In addition to intensive weekly mentoring (5 hours per week), every classroom in the in-service program evaluated by Grace et al. (2008) received classroom material (curriculum 2,000 USD and books 1,500 USD) and approximately 15,000 USD financial funds to independently select curriculum materials and classroom equipment (pp. 58-59). In contrast to this huge monetary support, the teachers in the consultation model by Palsha and Wesley (1998) received up to 200 USD to support environmental improvements (p. 246) and still made great gains on the ERS ($g = 0.86$). The intensive professional development program evaluated by Brinks (2007) was part of the EARLY Grant, in which tuition costs not covered by employers were paid by grant and teachers received a stipend at an hourly rate for attending training (pp. 45-46). These findings raise questions which role the provision of financial funds plays in the process of quality improvement and what strategies and support are needed to compensate release time of teachers.

Taking a closer look at *didactical principles*, it became apparent that teacher reflection, individualized approaches, and performance-based training are very common. A huge aggregated effect size ($g = 3.03$) was found for the 8-week course Best Practice in Prekindergarten Curriculum (Breffni, 2011). According to the author, the course was aimed explicitly at direct skill training with field experience, modeling and reflection, allowing teachers to apply new knowledge in classrooms during the training phase. Further, training was sustained over

time, interactive and collaborative, as well as grounded in practice (p. 179). The 16-month Early Childhood Leadership Training Program was based on the “assumption that immediate application of new learning to real-life situations reinforces what is learned” (Bloom & Sheerer, 1992, p. 584). The problem-centered and site-specific course, where teachers were actively engaged with 77 units (4 hours long), resulted in large effects ($g = 1.01$). Ongoing, intensive, individualized professional development based on individual needs was provided by the Professional Development Model (Brinks, 2007, pp. 44-45). The model consists of several training formats (i.e., coaching, network meetings, library with material, newsletters, workshops etc.). Coaching was a key component and based on the approach by Joyce and Showers (1981) to support teachers in their learning to ensure transfer to classrooms. This transfer is accomplished by teacher reflection, goal setting, identification of desired outcomes, identification of strategies, creation of an action plan, selection of coaching strategies, and implementation (cf. Brinks, 2007, p. 48). The intensive in-service approach by Cunningham (2007) with a 3-credit course and intensive weekly coaching (2 hours for 32 weeks) led to great effect sizes ($g = 0.86$), while the treatment without coaching did not do that well ($g = 0.38$). Professional development provided within the Project Great Start was content-focused and core competencies were matched to learning standards and the curriculum. Coaching was highly collaborative, interactive, individualized and facilitated reflection (Cunningham, 2007, pp. 79-81). In addition to acquiring knowledge through video reflection, discussion and lectures, the professional development from the LEEP Program focused on the implementation of new skills in classrooms through performance-based assignments (Dickinson & Caswell, 2007). In the 3-year professional development program by Grace et al. (2008), focusing on individual needs of teachers, teachers were supported through modeling and received written and oral feedback on their daily practice through mentors as well as feedback and individualized literacy coaching based on their ELLCO scores. The 15-week consultation model with weekly classroom observations and feedback, modeling and discussion of material provided to teachers by Howe et al. (2011) resulted in a remarkable impact ($g = 1.01$). The ExCELL-Program (Wasik & Hindman, 2011) consists of institutes, group training sessions and onsite training cycles as well as the provision of materials and led to significant quality improvements ($g = 1.29$). Within the training cycles, coaches provided training input, in-class modeling of target behavior, and in-class observation and feedback to teachers (p. 458). In the individualized consultation model by Palsha and Wesley (1998), teachers and consultants were trained to administrate the ERS. After training, both established a quality profile, discussed it and made individual technical assistant plans (p. 247). Within the Model of Competence Development by Sheridan (2001), the internal and external perspective of quality and reflection were major concerns of training. The scale-based training, with external ratings as a foundation, focused on reflections on the current state of quali-

ty, desired changes and the competence enhancement needed to achieve goals (p. 14). Another individualized professional development program was carried out by Powell et al. (2010). Teachers received a 2-day workshop with an emphasis on demonstration and guidance as well as bi-weekly coaching sessions with individual feedback to implement specific literacy-oriented practice over a period of 15 weeks (p. 303). In line with the guidelines from Buysse and colleagues (2009) and the integrative framework described in this thesis, the most effective in-service programs ($g > 0.80$) were ongoing, intensive and individualized to meet the participants' training needs. The training set individual priority and action plans, focused on reflection, provided guidance through modeling and videos of target behavior and offered oral or written feedback to teachers.

Focusing on programs with *video feedback*, the short-term program by Englund (2010) showed a substantial aggregated effect size ($g = 2.72$). The training focus was the CLASS dimension "language modeling". Teachers were video recorded and the researcher and teachers met 4 days in a row. During the meetings, a structured question-and-answer process with scripts was used to discuss the video tapes. The improvement of language-modeling strategies is an awareness process in which teachers review their practice through videos and come to understand the components of language modeling. Further, teachers passed through reflection and challenge phases in which they built strategies to improve their skills (Englund, 2010, p. 159). The 45-hour course from the LEEP-Program emphasizes teacher knowledge and reflection through the analysis of videotapes of best practice classroom activities and work examples (Dickinson & Caswell, 2007, p. 246). Within the remote coaching condition by Powell et al. (2010), teachers submitted videotapes of targeted instructional practice and received written feedback with direct links to video examples (p. 303). Comparing the remote coaching condition with video feedback with an onsite coaching condition, no significant contrast emerged, except for code-focused instruction (pp. 307-308). So far, findings suggest that videos provide equally effective opportunities compared to costly in-class modeling that initiates teacher reflection and provides feedback and guidance on target behavior.

4.2. Results on the impact of language and literacy training

4.2.1. Description of primary studies on language and literacy training

The meta-analysis on language and literacy development is a data subset of the systematic review project on the impact of in-service training on quality ratings and child outcomes. Overall, 48 studies with 59 different in-service treatments reported appropriate statistical data for conducting a meta-analysis (see table 19). More than one third of the studies ($n = 18$) were unpublished (e.g., research reports, conference papers and dissertations) and approximately 2 thirds were published studies ($n = 30$). Most of the studies were conducted in the United States ($n = 38$), except 5 from Germany, 4 from Canada and one from Australia. More than half of the studies included in the meta-analysis ($n = 26$) were published recently (after 2007). Overall, 25 studies used random assignment. Of these, 5 studies were randomized control trials (RCT) with assignment of individual teachers and 20 studies were cluster randomized trials (CRT) with assignment of classrooms or centers. Additionally, 21 studies applied a quasi-experimental design, with no specific random assignment ($n = 15$) or specific matching procedures ($n = 1$). Further, 7 non-experimental studies reporting the progress of a treatment group were included in the meta-analysis. Overall, data from 2,418 early childhood educators and 18,679 children were integrated in the meta-analysis. Large variations were found in sample sizes of teachers (range: 2 to 262) and of children (range: 24 to 1,786) at the beginning of the studies. The in-service programs varied between short-term programs with 3-day workshops (Girolametto et al., 2007; Rasberry, 2004) up to intensive 36-month programs (Shidler, 2009). However, 2 thirds had a maximum length of 12 months. Most of the programs ($n = 18$) run a regular kindergarten year with approximately 9 months of in-service training. Even though not all the studies explicitly reported the amount of professional development, short-term programs with a maximum of 45 training hours were the most common ($n = 26$). Nevertheless, a huge distribution exists (1.5 hours up to 506 hours). Again, the study from Shidler (2009) investigated different coaching formats ranging from 227 to 509 hours of individual coaching per classroom and provided data for several highly intensive treatments. Estimating a mean effect size for all language and literacy outcomes per treatment, negative and positive results appeared. Aggregated in-service training impacts on language scores ranged from $g = -0.14$ to 1.60 and the effects on literacy scores ranged from $g = -0.62$ to 2.21.

Analyzing study variance with CMA (Borenstein et al., 2005), the heterogeneity test showed some between-study variance ($Q = 1457$; $df = 58$; $p < .000$; $T^2 = .104$; $SE = 0.033$). Overall, 96 % of the observed variance across studies were due to systematic variance

($I^2 = 96.02$). The results indicated the use of a random effect model and a meta-regression to explain study variance. Investigating the stability of the summary effect, the Orwin's Fail Safe N test showed that 15 studies with a mean effect size of zero were needed to decrease the summary effect size to a negligible impact of $g < 0.20$. Classic Fail Safe N analysis pointed out that 4,651 studies with no results are needed to decrease the aggregated effect to $g = 0$. A test of sensitivity demonstrated a robust summary effect that slightly varied between $g = 0.33$ and $g = 0.39$, when one study was removed from meta-analysis.

Investigating publication bias, a chi-square analysis indicated that effect size variance was not significantly related to sample size of teachers ($\chi^2 = 10026$; $df = 9966$; $p = .333$), sample size of children ($\chi^2 = 15534$; $df = 15402$; $p = .224$), publication type ($\chi^2 = 305$; $df = 301$; $p = .425$) or standard error ($\chi^2 = 91800$; $df = 91506$; $p = .246$). Egger regression ($Intercept = 2.62$; $SE = 1.10$; $p = .021$) indicated some asymmetry, where the study estimate is related to the size of the study.

Aggregating the effect sizes per treatment for literacy and language scores, large variations in effect size were found. The aggregated impact per in-service treatment on the aggregation of language scores varied from $g = -0.10$ to 1.60 and for the aggregation of literacy values from $g = -0.62$ to 2.21 per professional development program (see table 19).

Altogether, 306 effect sizes with a range from $g = -1.97$ to 3.95 could be extracted from (quasi)-experimental studies and non-experimental studies. Effect sizes are based on the contrast between experimental and control groups at posttest or from change scores from pre- to posttest. Of the 306 effect sizes (see figure 7), 149 describe the impact on language outcomes (range = -1.79 to 3.95) and 157 on literacy scores (range = -1.54 to 2.76). Approximately 40 % of effect sizes on language ($k = 61$) or literacy development ($k = 56$) display no or negative in-service impacts ($g < 0.20$), according to Cohen (1988).

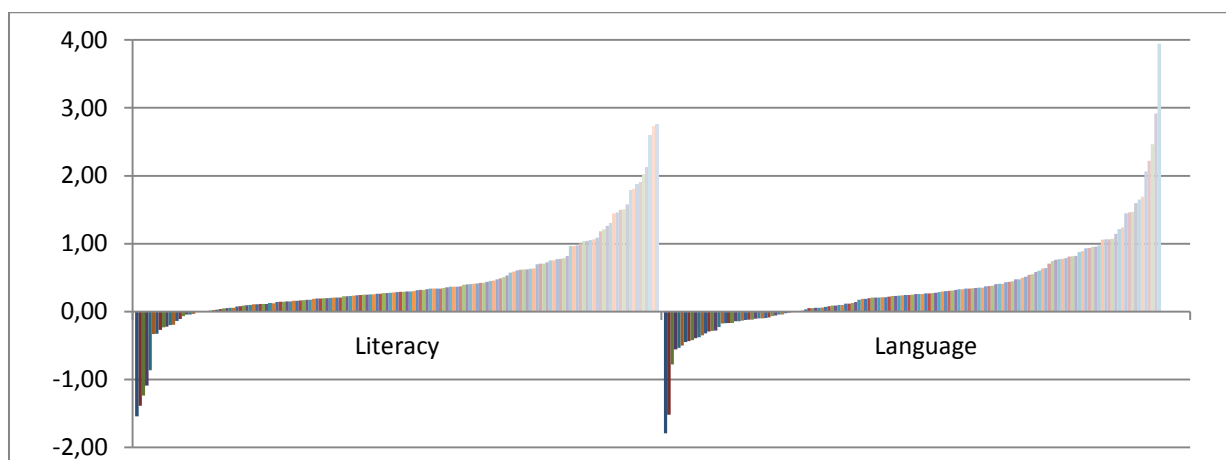


Figure 7. Effect sizes distribution for language and literacy outcomes

Table 19. Overview of studies included in meta-analysis on language and literacy

Author	Year	Country	Pubtype	Design	Sample teacher	Sample children	ES language	ES literacy	WS	CS	ONL	COP	ON	Duration (months)	Amount (hours)
Ahrens	2009	USA	unpub	CRT	6	24	0.23						x	2.50	1.50
Assel et al. (a)	2007	USA	pub	CRT	76	603	0.09		x				x	9.00	59.00
Assel et al. (b)	2007	USA	pub	CRT	76	603	0.14		x				x	9.00	36.50
Baker & Smith	1999	USA	pub	QNE	17	186		-0.34				x		12.00	n.r.
Barnett et al.	2008	USA	pub	CRT	18	218	0.17	-0.04	x			x	x	9.00	59.00
Beller & Beller	2009	Germany	unpub	RCT	38	164	0.08		x				x	7.00	81.00
Beller, Merckens, & Preissing	2007	Germany	unpub	QNE	31	155	0.25		x				x	6.00	81.00
Bos et al.	1999	USA	pub	QNE	31	324	0.63	0.49		x			x	9.00	n.r.
Brown	2008	USA	unpub	QNE	3	46	0.41	1.54		x			x	4.00	20.00
Bryant et al.	2009	USA	unpub	RCT	108	236	0.37						x	9.00	30.00
Buschmann et al.	2010b	Germany	pub	QNE	30	33	0.94			x				7.00	18.75
Buyse, Castro & Peisner-Feinberg	2010	USA	pub	CRT	55	193	-0.05	0.16	x			x	x	9.00	40.00
Clancy Menchetti	2006	USA	unpub	CRT	40	500	-0.01	0.08	x	x			x	24.00	216.00
Cordell	2010	USA	unpub	OGPP	3	34	1.60			x			x	3.00	n.r.
Cusumano et al. (a)	2006	USA	pub	QNE	41	386	-0.08	0.16		x			x	4.00	35.81
Cusumano et al. (b)	2006	USA	pub	QNE	41	386	-0.14	0.15	x					4.00	30.00
Domitrovich, Cortes, & Greenberg	2007	USA	pub	QNE	20	275	0.14		x				x	9.00	28.50
Downer et al. (a)	2011b	USA	pub	CRT	242	1338	-0.04	0.14	x		x		x	21.00	n.r.
Downer et al. (b)	2011b	USA	pub	CRT	242	1338	0.06	0.21	x		x			21.00	n.r.
Flowers et al.	2007	Canada	pub	CRT	16	64	0.52			x			x	3.50	23.50
Gallagher et al.	2011	USA	pub	RCT	16	48	0.20	0.16					x	9.00	14.50
Girolametto, Weitzman, & Greenberg	2003	Canada	pub	CRT	16	64	1.25			x			x	3.50	23.50
Girolametto, Weitzman, & Greenberg	2004	Canada	pub	CRT	17	68	-0.10			x			x	1.50	9.25
Girolametto et al.	2007	Canada	pub	RCT	16	64	0.57		x					0.10	12.00
Graul & Zeece	1990	USA	pub	OGPP	18	28	0.47			x				1.75	14.00
Helmer et al.	2011	Australia	pub	OGPP	11	157		0.83	x		x		x	9.00	n.r.
Hsieh	2005	USA	unpub	OGPP	5	58	0.29	0.75					x	2.50	13.50
Kopacsi & Hochwald	1998	USA	unpub	QNE	21	1219		0.46	x				x	12.00	n.r.
Krause, Kofler, & Hofbauer	2011	Germany	unpub	QNE	118	418	-0.06	-0.62	x				x	12.00	115.00
Landry et al.	2009	USA	pub	CRT	262	1786	0.19	0.20			x		x	9.00	72.00
Landry, Swank, & Anthony	2011	USA	pub	CRT	214	1571	0.12	0.10			x			13.50	54.00

(Table 19. continued)

Author	Year	Country	Pubtype	Design	Sample teacher	Sample children	ES language	ES literacy	WS	CS	ONL	COP	ON	Duration (months)	Amount (hours)
Lee	2010	USA	unpub	CRT	25	383	1.60			x	x			9.00	n.r.
Lonigan et al. (a)	2011	USA	pub	CRT	51	808	0.32	0.15	x				x	9.00	68.00
Lonigan et al. (b)	2011	USA	pub	CRT	51	808	0.20	0.16	x					9.00	32.00
Mashburn et al.	2010	USA	pub	CRT	182	1165	0.27		x		x		x	21.00	43.00
McCutchen et al.	2002	USA	pub	QWM	44	492		0.40	x				x	9.50	70.00
McGill-Franzen et al.	1999	USA	pub	CRT	18	429	0.98	2.21	x	x				4.00	30.00
Mohler et al.	2009	USA	pub	QNE	24	768		0.31	x	x			x	24.00	n.r.
Morris et al.	2010	USA	unpub	QNE	52	623	0.27		x	x			x	9.00	162.00
O'Connor (a)	1999	USA	pub	QNE	10	154	0.42	0.88					x	8.00	n.r.
O'Connor (b)	1999	USA	pub	QNE	17	311	-0.01	0.55	x					6.00	12.00
Onchwari & Keengwe	2010	USA	pub	QNE	44	626	0.25	0.20					x	5.00	40.00
Podhajski & Nathan	2005	USA	pub	QNE	86	126		0.34	x				x	6.00	21.00
Powell et al.	2010	USA	pub	RCT	88	759	0.12	0.22	x				x	4.00	37.00
Raspberry	2004	USA	unpub	CRT	4	38		2.03	x					0.10	n.r.
Rosof	2006	USA	unpub	OGPP	66	192		0.05	x			x	x	24.00	n.r.
Rudd	2003	USA	unpub	CRT	59	121	0.03		x				x	0.17	n.r.
Schröder et al.	2011	Germany	unpub	OGPP	20	49	0.31		x				x	6.00	n.r.
Shidler (a)	2009	USA	pub	OGPP	2	43	0.38			x			x	36.00	269.00
Shidler (b)	2009	USA	pub	OGPP	2	42	0.30			x			x	36.00	266.00
Shidler (c)	2009	USA	pub	OGPP	2	45	0.79			x			x	36.00	322.00
Shidler (d)	2009	USA	pub	OGPP	2	39	0.09			x			x	36.00	252.00
Shidler (e)	2009	USA	pub	OGPP	2	44	0.60			x			x	36.00	249.00
Shidler (f)	2009	USA	pub	OGPP	2	54	0.38			x			x	36.00	227.00
Shidler (g)	2009	USA	pub	OGPP	2	39	0.77			x			x	36.00	509.00
Tyler	2009	USA	unpub	QNE	4	170		0.23					x	9.00	n.r.
Vernon-Feagans et al.	2010	USA	unpub	CRT	75	151	0.16	0.46			x		x	9.00	18.00
Wasik & Hindman	2011	USA	pub	CRT	30	541	0.21	0.17	x	x			x	9.00	43.75
Wasik, Bond & Hindman	2006	USA	pub	CRT	16	207	0.61	-0.33		x			x	9.00	36.00

Note. Pubtype=publication type; pub=published in journals or as book chapter; unpub=unpublished paper e.g., technical paper, dissertation or governmental report; design=study design; RCT=randomized control trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center to groups; QWM=quasi-experimental design with matching procedure/equivalent groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design; OGPP=one group pre-post measure; ES language=aggregated effect size of treatment on language scores; ES literacy=aggregated effect size of treatment on literacy scores; WS=workshop; CS=course and in-service with ongoing sessions; ONL=online training or blended learning; COP=community of learning or community of practice; ON=onsite support like coaching, consulting etc.

4.2.2. Meta-analysis and moderator analysis on language and literacy

A small significant weighted summary effect size ($g' = 0.38$; $SE = 0.045$; $k = 306$; $CI = 0.29$ to 0.47 ; $Z = 8.376$; $p < .000$) could be estimated for the impact of in-service programs on language and literacy scores, including 48 studies with 59 treatments. The confidence interval of the weighted summary effect is relatively narrow in the small effect size range. Studies were weighted in meta-analytic estimates according to their variance, ranging from 0.07 to 1.97 %. Aggregating the impact per treatment, the synthesized effect sizes ranged from $g = -0.32$ to 2.03 (see table 20). The effect size estimates of 3 in-service training programs showed negative effects on language and literacy development. Confidence intervals from the aggregated effects per treatment were relatively wide indicating an imprecise intervention effect with different degrees of benefit. The lower confidence interval estimates from 18 treatments were in the negative range.

Meta-analytic estimations revealed a small aggregated summary effect size ($g = 0.40$; $SE = 0.054$; $k = 306$; $RIGLS = 777.6$; $\Delta\chi^2(58) = 2769$; $p < .000$) for the impact of in-service programs on language and literacy scores. Differential analysis exposed a small significant impact on language scores ($g = 0.36$; $SE = 0.056$) and a slightly higher small significant effect on literacy scores ($g = 0.45$; $SE = 0.056$) with tendency to medium. Compartmentalizing language and literacy development into more specific domains, differential effects also appeared.

Findings from the multi-level moderator analysis showed that the impact of in-service programs on receptive vocabulary ($g = 0.23$) and non-word repetition skills ($g = 0.30$) was considerably smaller than for other subdomains (see table 21). In contrast, a significant medium size professional development effect was achieved for children's comprehension abilities ($g = 0.56$). Emphasizing literacy-specific categories, significant small effects with tendency to medium were found for children's elision abilities ($g = 0.48$) and concepts about print ($g = 0.49$). A medium size in-service effect appeared for children's writing abilities ($g = 0.64$). However, professional development affected phonological awareness scores considerably less ($g = 0.25$) than other domains.

Model 2 examined the significance of outcome domain as explanatory moderators adjusting for methodological effect modifiers. In particular, results from model 1 (see table 22) showed that effect sizes were significant predicted by type of contrast (1=contrast between experimental and control groups at posttest), differences at pretest (control group with higher values) and score level (1=total; 0=subcategories). The 3 covariates were entered in the equation simultaneously and all 3 remained as significant predictors. The correlations between those methodological modifiers were negligible ($r < .09$).

Table 20. Meta-analysis on language and literacy development

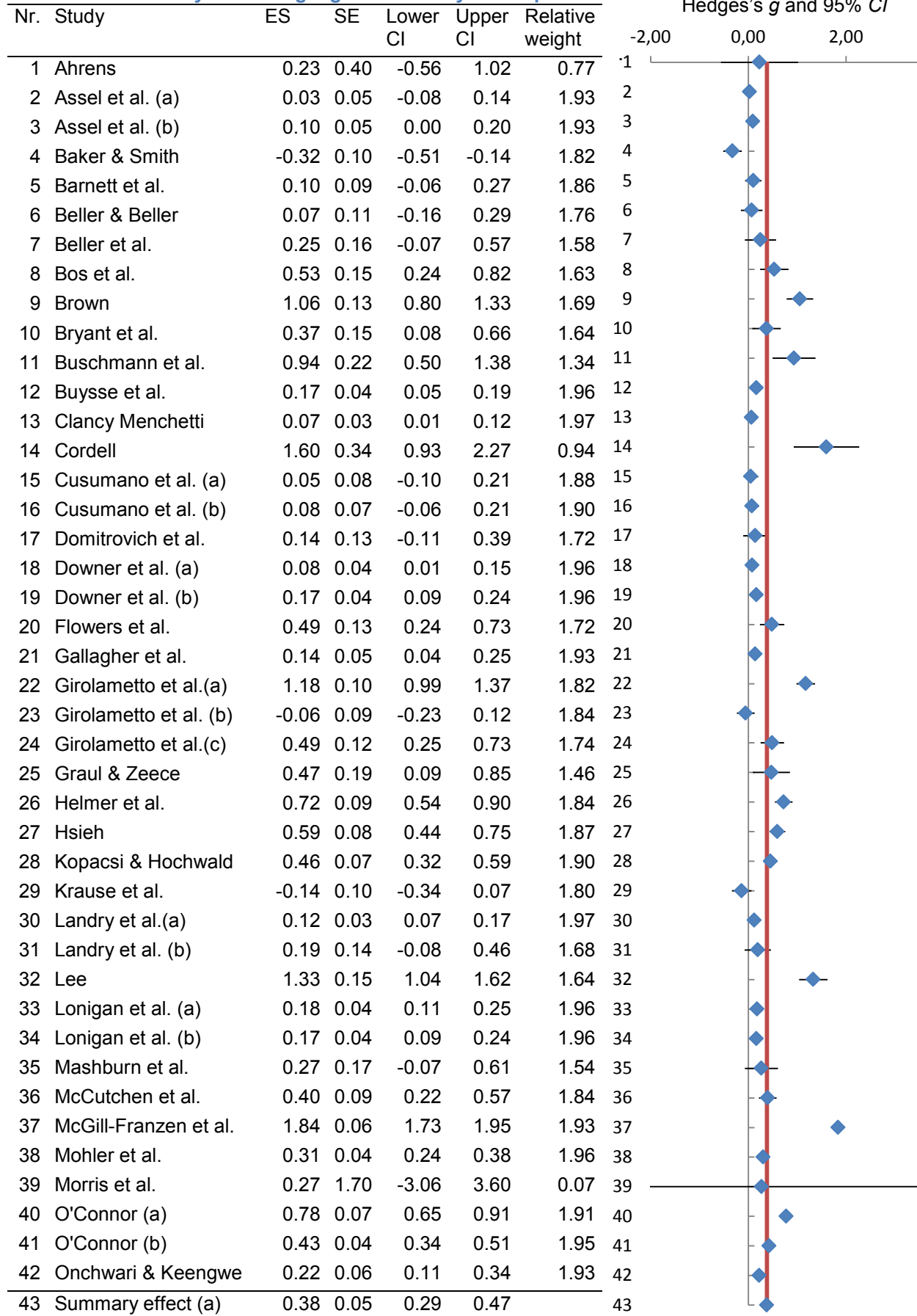
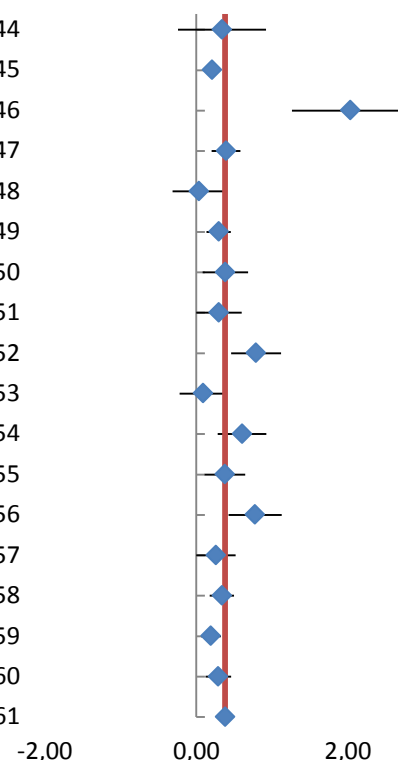


Table 20. (continued)

Nr.	Study	ES	SE	Lower CI	Upper CI	Relative weight
44	Podhajski & Nathan	0.34	0.30	-0.24	0.92	1.08
45	Powell et al.	0.21	0.03	0.14	0.27	1.97
46	Raspberry	2.03	0.39	1.26	2.80	0.80
47	Rosof	0.39	0.10	0.21	0.58	1.83
48	Rudd	0.04	0.18	-0.32	0.39	1.51
49	Schröder et al.	0.30	0.08	0.14	0.46	1.87
50	Shidler (a)	0.38	0.16	0.08	0.69	1.61
51	Shidler (b)	0.30	0.16	-0.01	0.60	1.61
52	Shidler (c)	0.79	0.17	0.46	1.12	1.56
53	Shidler (d)	0.09	0.16	-0.22	0.40	1.60
54	Shidler (e)	0.60	0.16	0.29	0.92	1.59
55	Shidler (f)	0.38	0.14	0.10	0.65	1.67
56	Shidler (g)	0.78	0.18	0.42	1.13	1.51
57	Tyler	0.26	0.13	0.00	0.52	1.70
58	Vernon-Feagans et al.	0.34	0.08	0.18	0.49	1.87
59	Wasik & Hindman	0.19	0.07	0.06	0.32	1.90
60	Wasik et al.	0.29	0.09	0.12	0.46	1.85
61	Summary effect (b)	0.38	0.05	0.29	0.47	

Hedges's g and 95% CI

In model 2, the explorative theoretical modifier was entered simultaneously in the equation with the 3 mentioned methodological variables. Findings from model 2 confirmed the findings from the unadjusted regression (model 1 in table 21) on receptive vocabulary, language comprehension, non-word repetition, concepts about print and writing abilities. However, findings on elision abilities and phonological awareness were not significantly retained when they were adjusted for methodological artefacts. After controlling for 3 methodological variables, the effect sizes for segmentation and beginning sound abilities were larger and the effect size for alphabetic knowledge smaller.

In table 22, theoretical and methodological effect modifiers were tested through a random multi-level procedure individually (model 1). Several methodological variables are known to modify effect sizes from different meta-analyses (cf. Blok et al., 2005; Blume et al., 2010; Wilson & Lipsey, 2001). Indeed, the comparison group type emerged as a significant modifier in this data set. Effect sizes estimated from mean contrasts between experimental and control groups at posttest were significantly smaller than pre-post gains or contrasts with alternative treatments. Further, differences in mean scores at pretest influenced the effect size considerably. In addition, effect sizes estimated from total scores were smaller ($g = 0.34$) than those extracted from subdomains ($g = 0.44$). The informant did not modify the effect size.

In model 2, theoretical effect modifiers were entered in addition to the 3 covariates (control group score, total score, and contras) to analyze the independent influence of the exploratory variable. In addition to the 3 methodological covariates from model 2, outcome domain

(1=language) was used as fourth covariate in the equation in model 3. The methodological variables correlated only modestly with outcome domain ($r < .23$).

As theoretical effect modifiers, *training intensity and duration* were examined. The regression coefficients for training amounts in hours (*regression coefficient* = .000; SE = 0.001) and duration in months (*regression coefficient* = -.004; SE = 0.005) did not support the assumption that in general more intense training is better than less intense programs. Splitting amount and duration into theoretical categories related to university term and credits, professional development programs with 15 to 30 training hours were the most effective category ($g = 0.69$). The findings on training amount remained, adjusting for methodological covariates and outcome domain. Supplementary to these results, professional development that is carried out between 3 to 6 months seems to be the most effective duration category, when methodological artefacts and outcome domains were used as covariates.

Table 21. Moderator analysis of in-service training effects on language and literacy domains

Characteristic	Model 1				Model 2	
	$k=306$				$k=306$	
	k	g	Residual variance	$\Delta\chi^2$	Significance of moderator	Residual variance
Language and literacy	306	0.40*	.152	2886*	sig.	.145
Language	148	0.36*	.155	967*	sig.	.145
Language composite	26	0.44	.154	90*	n.s.	.147
Receptive language	29	0.23*	.161	154*	sig.	.143
Expressive language	31	0.37	.152	208*	n.s.	.145
Utterances	35	0.68	.146	376*	n.s.	.140
Comprehension	27	0.56*	.154	120*	sig.	.147
Non-word repetition	8	0.30*	.151	192*	sig.	.144
Literacy	157	0.45*	.155	1939*	sig.	.145
Literacy composite	13	0.39	.152	122*	n.s.	.146
Alphabet knowledge	26	0.37	.152	198*	sig.	.150
Beginning sound	9	0.43	.152	45	sig.	.145
Blending	14	0.42	.152	117*	n.s.	.146
Elision	8	0.48*	.153	63	n.s.	.146
Rhyming	17	0.38	.152	64	n.s.	.146
Phonemic segmentation	13	0.62	.159	185*	sig.	.153
Phonological awareness	1	0.25*	.152	40	n.s.	.145
Phonemic awareness	4	0.40	.152	24	n.s.	.146
Early reading	14	0.44	.152	44	n.s.	.146
Dictation	6	0.53	.151	19	n.s.	.144
Concepts about print	13	0.49*	.151	325*	sig.	.145
Book knowledge	6	0.32	.152	56	n.s.	.145
Writing	5	0.64*	.148	443*	sig.	.141

Note. * $p < .05$; sig=significant with $p < .05$; n.s.=not significant; Model 1=unadjusted; Model 2=adjusted for methodological covariates (contrast, control group had higher values at pretest and total score).

Looking at the *delivery format* of professional development (see table 22), at first glance, a significant finding emerged for traditional training formats. While the delivery of workshops or coursework only are in line with the aggregated impacts on language and literacy scores, the

combination of workshop and course is significantly more effective ($g = 1.84$) than other professional development programs ($g = 0.36$). The highly effective professional development program ($g = 1.84$) by McGill-Franzen and colleagues (1999), in which teachers received approximately 30 hours of training and books for the classroom, used this format only. Exploratory analysis showed that the effect disappeared when the study by McGill-Franzen et al. (1999) was excluded, and alternative treatments using courses and onsite support became the most promising approach (*regression coefficient* = .224; SE = 0.099; $k = 300$).

With respect to *modern learning formats*, online learning and web-based approaches ($n = 8$) as well as programs ($n = 6$) that used onsite support solely (e.g., coaching, consulting or mentoring) are equally effective in comparison to other types. The same tendency emerged in the moderator analysis for the mixture of delivery methods ($n = 41$) which combined traditional formats or online with onsite support. A variety of delivery formats resulted in a smaller effect size ($g = 0.22$). Taking a closer look at different combinations, 2 (non-significant) trends appeared. At least the combination of coursework and onsite support seems to be positively related to the magnitude of the training effect (*regression coefficient* = .187; SE = 0.124), while the combination of workshop and onsite support gave a negative association (*regression coefficient* = -.245; SE = 0.130). Moderator analysis showed that findings on delivery persisted adjusting for methodological artefacts (model 2) and outcome (model 3).

Addressing the question if *children from different backgrounds* profit equally from professional development, moderator analysis showed that in-service training was more effective when the majority of the child sample was described as being at risk ($g = 0.53$) in comparison to studies with non-disadvantaged children ($g = 0.15$). This finding remained significant after adjusting for covariates in models 2 and 3. In summary, disadvantaged children benefited notably in their language and literacy development when their teachers received training.

The research question if the *fidelity of the implemented strategies or curriculum* moderates in-service training effectiveness could be examined with a limited data set. In 23 studies (29 treatments), a curriculum or specific instructional strategies/activities were implemented. Moderator analysis did not support the assumption that curriculum implementation is a significant effect modifier, after testing whether curriculum implementation was a major target (*regression coefficient* = -.169; SE = 0.278, $k = 306$) or curriculum/specific activities were part of the training (*regression coefficient* = -.030; SE = 0.156; $k = 306$). However, only 9 studies (12 treatments) provided information on the fidelity of curriculum or strategy implementation (i.e., Assel et al., 2007; Barnett et al., 2008; Brown, 2008; Clancy-Menchetti, 2006; Cusumano et al., 2006; Hsieh, 2005; Mohler et al., 2009; Vernon-Feagans et al., 2010; Wasik & Hindman, 2011). Exploratory analysis did not provide evidence (*regression coefficient* = .186; SE = 0.207; $k = 91$) that in-service training effects are statistically moderated by high implementation fidelity (fidelity > 80 %).

Table 22. Moderator analysis of in-service effects on language and literacy

Characteristic	k	Model 1 k=306				Model 2 k=306		Model 3 k=306	
		g	Residual variance	$\Delta\chi^2$		Significance of moderator	Residual variance	Significance of moderator	Residual variance
Publication type				2306*					
Published	229	0.36	.153		n.s.	.148	n.s.	.148	
Unpublished	77	0.49*	.153		n.s.	.148	n.s.	.148	
Methodology									
Comparison (EG vs. CG)				2360*	-	-	-	-	-
Yes	251	0.33*	.143		-	-	-	-	-
No	55	0.58*	.143		-	-	-	-	-
Difference at pretest					-	-	-	-	-
Experimental group higher	128	0.53*	.158	1522*	-	-	-	-	-
Control group higher	110	0.25*	.149	816*	-	-	-	-	-
Total score				895*	-	-	-	-	-
Yes	119	0.34*	.156		-	-	-	-	-
No	187	0.44*	.156		-	-	-	-	-
Informant									
Child (test)	228	0.38	.150	2243*	n.s.	.143	n.s.	.142	
Teacher	22	0.34	.152	144*	n.s.	.146	n.s.	.146	
Observer	37	0.58	.149	384*	n.s.	.141	n.s.	.138	
Parents	3	0.94	.150	6	n.s.	.146	n.s.	.145	
In-service training intensity									
Duration									
Less than 3 months	22	0.57	.154	261*	n.s.	.148	n.s.	.148	
3 to < 6 months	65	0.54	.147	1136*	n.s.	.136	sig.	.136	
6 to < 9 months	147	0.37	.155	688*	n.s.	.148	n.s.	.148	
9 to < 12 months	24	0.14	.149	187*	n.s.	.142	n.s.	.141	
12 to < 24 months	41	0.20	.149	337*	n.s.	.138	n.s.	.137	
24 to 36 months	29	0.52	.152	277*	n.s.	.149	n.s.	.148	
Amount									
Less than 15 hours	48	0.33	.155	318*	n.s.	.148	n.s.	.148	
15 to < 30 hours	49	0.69*	.131	1012*	sig.	.123	sig.	.120	
30 to < 45 hours	59	0.40	.153	337*	n.s.	.147	n.s.	.146	
45 to < 70 hours	35	0.31	.149	313*	n.s.	.144	n.s.	.144	
70 to < 100 hours	6	0.17	.152	14	n.s.	.148	n.s.	.148	
100 to < 200 hours	16	0.03	.152	60	n.s.	.146	n.s.	.146	
> 200 hours	14	0.41	.155	189*	n.s.	.145	n.s.	.145	

(Table 22 continued)

Characteristic	k	Model 1			Model 2		Model 3	
		k=306 g	Residual variance	$\Delta\chi^2$	Significance of moderator	Residual Variance	Significance of moderator	Residual variance
<u>Traditional delivery formats</u>								
Course only				6	n.s.	.147	n.s.	.147
Yes	4	0.70	.152					
No	302	0.39*	.152					
Workshop only				194*	n.s.	.149	n.s.	.149
Yes	29	0.49	.156					
No	277	0.39*	.156					
Workshop & course				808*	sig.	.093	sig.	.094
Yes	6	1.84*	.101					
No	300	0.36*	.101					
<u>Innovative delivery formats</u>								
Web-based programs				350*	n.s.	.149	n.s.	.149
Yes	54	0.39	.155					
No	252	0.40*	.155					
Only onsite				169*	n.s.	.148	n.s.	.149
Yes	39	0.38	.155					
No	267	0.40*	.155					
<u>Combined learning formats</u>								
Mixed methods				1320*	sig.	.148	sig.	.148
Yes	197	0.22*	.172					
No	119	0.73*	.172					
Workshop & onsite				397*	n.s.	.145	n.s.	.145
Yes	77	0.20	.146					
No	229	0.45*	.146					
Course & onsite				395*	n.s.	.147	n.s.	.146
Yes	49	0.54	.148					
No	257	0.35*	.148					
Setting				2209*	sig.	.172	sig.	.172
Majority of child sample at risk	228	0.53*	.201					
Non-disadvantaged children	78	0.15*	.201					

Note. * $p < .05$; $\Delta\chi^2$ =chi-square test of residual variance n.s.=not significant; sig.=significant; Model 1=unadjusted; Model 2=significance-tested with methodological covariates (control group score higher at pretest, total score, contrast EG vs. CG at posttest); Model 3=significance-tested with methodological covariates (see model 2) and outcome domain (literacy or language score).

4.2.3. Narrative review of effective studies on language and literacy

Taking a closer look at highly effective professional development programs (defined as programs with an aggregated effect of $g > 0.80$), it is obvious that professional development *providers are highly skilled and knowledgeable* of the content they deliver (see table 23). In particular, in the programs by Buschmann et al. (2010b) and Girolametto et al. (2003) trainers pass through a certification procedure with a “train the trainer” curriculum. In other studies, researchers (authors, graduate student etc.) provided the training to teachers. However, none of the highly effective professional developments is a large-scale program. Hiring (external) coaches for large-scale in-service programs may provide an unpredictable risk for losing effectiveness. Providing intensive preparation for trainers with a credential procedure seems to be necessary to ensure the quality of professional development.

Further, most of the highly effective professional development programs could be categorized as “*interaction training*” that qualifies educators to use *specific techniques to engage children* in individual, small or large group discussions. In particular, child center strategies (e.g., be face-o-face, observe-wait-listen, or follow child’s lead), specific interaction-promoting strategies (e.g., variety of questions or turn-taking), and language-modeling strategies (e.g., variety of language, extend or expand utterance) in early childhood classrooms were facilitated through different training formats. In particular, effective questioning, such as the use of open-ended questions, was often favored (Brown, 2008; Buschmann et al., 2010b; Girolametto et al., 2003; Lee, 2010). Target behavior was demonstrated by video sequences either through a short film or by online sequences (Lee, 2010; Rasberry, 2004). Further, teachers received video feedback for their implementation of the new strategies and techniques (Buschmann et al., 2010b; Girolametto et al., 2003).

Generally, most of the effective programs included *performance-based assignments* (specific exercises) or activity lists for teachers to practice the new skills in their classrooms. Most of these performance-based assignments were tailored for specific situations (book reading) with target teaching behavior. Teachers received tutorials to exercise specific teaching behavior in specific situations in their classrooms. Performance-based assignments also used daily activity plans that reminded teachers to exercise specific behavior in daily routines. Given such specific guidelines, the implementation of target behavior or techniques in classrooms seems to be more essential and to foster the retention of new skills. Additionally, providing individual feedback with or without video seems to be needful to guide teachers in implementing new interaction techniques in classrooms.

Table 23. Narrative review of effective in-service programs on language and literacy

Author	Trainer	Delivery format	Focus
Brown (2008)	Graduate student and university coach	- Course (4 x 1.25 hours) - Consultation 2 times a week for 15 weeks	- Implementation of literacy activities - Questioning techniques - Phonological awareness lessons
Buschmann et al.(2010b)	Certified trainer & author	- 5 full day course sessions including exercises for the classroom in book reading between training sessions - Video feedback (group)	Interaction-training with: - child-centered strategies - interaction-promoting strategies - language-promoting strategies
Girolametto, Weitzman & Greenberg (2003)	Certified trainer (speech language pathologist)	- 14 weeks program - Course (6 x 2,5h) - Exercises between course - Video feedback (individual)	- Implementation of “Learning Language and Loving it” with: - child-centered response - interaction-promoting strategies - language modeling response
Helmer et al. (2011)	Researcher (literacy coach)	- Introduction institute - Ongoing support by literacy coach - Teachers receive the web-based ABRACADABRA software tool	- Implementing the ABRACADABRA software tool that helps teachers to develop and structure literacy activities
Lee (2010)	MyTeachingPartner Staff	- Course - Online sessions - 7 demonstration videos	- MTP general curriculum support for MTP Science Curriculum - Specific questioning support (e.g., open ended questioning)
McGill-Franzen et al. (1999)	n.r.	- 3 full day sessions - 7 x 2-hour session - Exercises in book reading (log book). - Teachers receive books for the classroom and parents’ libraries	- Interactive techniques - Reading aloud - Literacy environment - Literacy activities - Small group lessons based on books read
O’Conner (intensive training) (1999)	Author	- Series of individual meetings	- Implementing phonological and print awareness activities from “Ladders to Literacy” in classrooms
Rasberry (2004)	Author	- Watching a “best practice” film - Discussion with author - Teachers receive a list of activities for daily practice	- Conversational techniques during book reading

Note. n.r.=not reported.

4.3. Results on the impact on social-emotional and behavioral skills

The impact of in-service programs on the promotion of social abilities and the reduction of challenging behavior is at the center of this section. The magnitude of the effects were estimated and theoretical and methodological effect modifiers were tested.

4.3.1. Description of primary studies on social-emotional and behavioral skills

Overall, 13 studies reported appropriate statistical data to conduct a meta-analysis (see table 24). Approximately 2 third of the studies were conducted in the United States ($n = 9$). Avoiding publication bias, a good ratio (1:2) of unpublished studies ($n = 5$), including research reports, conference papers and dissertations, and published studies ($n = 8$) could be found for the meta-analysis. All the studies, except one by Rhodes and Hennessey (2000), were carried out quite recently. This indicates that the quality of studies, data and statistical procedures may have improved over the last few years. Overall, the meta-analysis comprised of data from 538 teachers and 2,578 children. Significant variations were found in both samples, ranging from 3 to 108 early childhood educators and 13 to 623 preschool children participating at the beginning of the studies. The statistical data used in the meta-analysis were retrieved from group contrasts and data on changes between pretest and posttest.

Studies using workshops, courses, or onsite support as delivery formats were included in the meta-analytic procedure. Due to insufficient information, some studies investigating the impact of online resources (Driscoll et al., 2011) or community of practice formats (Barnett et al., 2008) on social and behavioral skills could not be included. The professional development duration ranged from 1.25 months to an academic year of 9 months. While 5 studies analyzed a short-term approach of 3 months or less, 5 studies examined intensive long-term formats which lasted a full academic year. Only 5 studies with 6 treatments had an amount of 30 or more training hours. The aggregated impact of teacher training per treatment on social skills ranged from $g = -0.24$ to $g = 2.39$. The impact of professional development on challenging behavior, aggregating all measures on external and internal problem behavior per treatment, varied from $g = -0.55$ to $g = 0.59$.

Overall, 86 effect sizes could be extracted from the 13 studies. Of these, 44 effect sizes included information on the impact of professional development programs on social-emotional skills, varying from $g = -0.22$ to $g = 2.87$ (see figure 8). To measure the effectiveness of professional development, different informants and instruments were used. Social-emotional growth was measured through direct child assessment (e.g., ACES Assessment of Chil-

dren's Emotional Scale, DPI Denham Puppet Interview), teacher ratings (e.g., PKBS Preschool and Kindergarten Behavior Scale, PBS Positive Behavior Scale, SCP Social Competence Performance Checklist), classroom observations of target child behavior or classroom behavior (e.g., CCOF Classroom Competence Observation Form), or parental ratings (e.g., HSCS Head Start Competence Scale, PKBS Preschool and Kindergarten Behavior Scale).

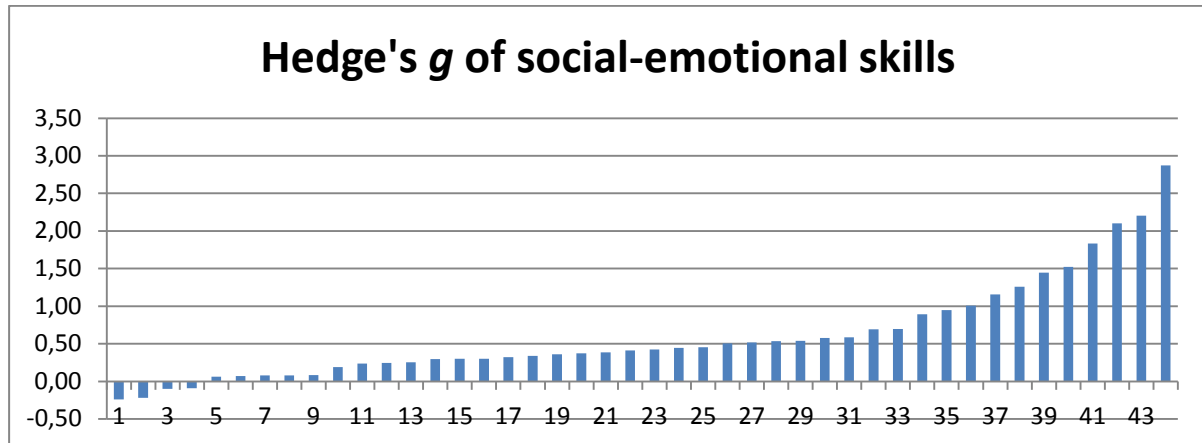


Figure 8. Effect size distribution on social-emotional outcomes

Forty-two effect sizes described the impact of in-service programs on reduced behavioral problems. The effects ranged from $g = -0.22$ to $g = 1.26$ (see figure 9). Behavioral problems were also measured through 4 different informants and with a variety of instruments. Contrast or change scores were obtained through child assessment scores (e.g., CST Challenging Situation Test), teacher ratings (e.g., PBI Problem Behavior Index, PBS Positive Behavior Scale, SCP Social Competence Performance Checklist), observations (e.g., CCOF Classroom Competence Observation Form, DPIC Dyadic Parent-Child Interaction Coding System) and parental reports (e.g., EDBI Eyberg Child Behavior Inventory, SDQ Strengths and Difficulties Questionnaire, PKBS Preschool and Kindergarten Behavior Scale).

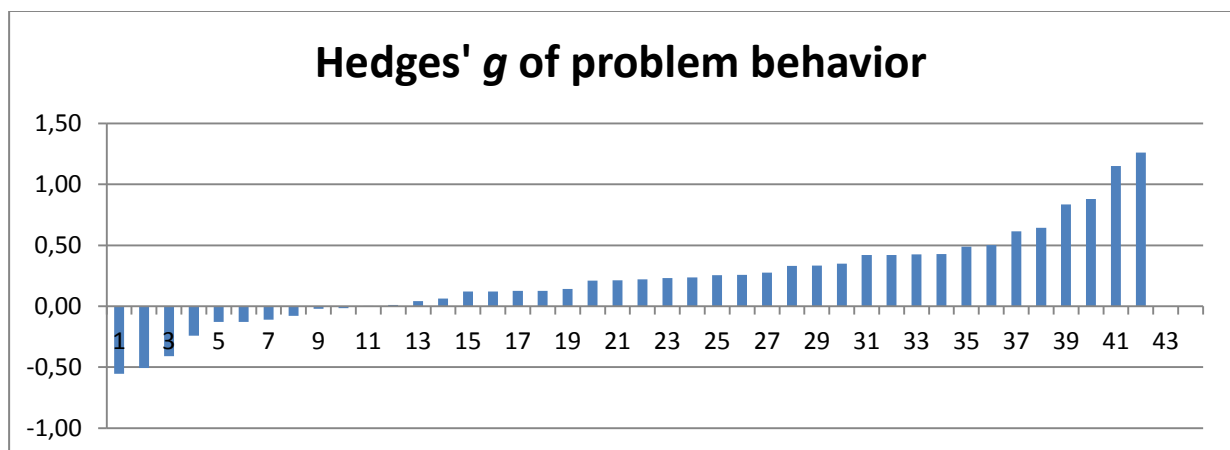


Figure 9. Effect size distribution of problem behavior outcomes

Table 24. Overview of studies included in meta-analysis on social and behavioral outcomes

Author	Year	Country	Pubtype	Sample teacher	Sample children	Design	ES social	ES problem	WS	Course	Onsite	Months	<30 h
Baker-Henningham et al.	2012	Jamaica	pub	73	225	experimental	0.22	0.32	x		x	5.00	
Bryant et al.	2009	USA	unpub	108	236	experimental	-0.24	0.01			x	9.00	
Cordell	2010	USA	unpub	3	34	one group	1.83			x	x	3.00	x
Domitrovich et al.	2007	USA	pub	20	275	quasi-experimental	0.35	0.19	x		x	9.00	x
Gallagher et al.	2011	USA	pub	16	48	experimental	0.53				x	9.00	x
Girard et al.	2011	Canada	pub	17	68	experimental	0.42	0.13	x		x	1.50	x
Lloyd & Millenky	2011	USA	unpub	91	600	quasi-experimental	0.08	-0.22	x	x	x	9.00	
Morris et al.	2010	USA	unpub	51	623	experimental	0.07	-0.07	x	x	x	9.00	
Rhodes & Hennessy	2000	UK	pub	33	66	quasi-experimental	0.60			x		6.00	
Shernoff & Kratochwill	2007	USA	pub	8	13	experimental		-0.24			x	1.25	x
Snyder et al.	2011	USA	pub	28	136	experimental	0.79	0.23		x	x	3.00	x
Stoiber & Gettinger	2011	USA	pub	70	90	experimental	1.15	0.59	x		x	4.50	x
Thanasetkorn	2009	Thailand	unpub	20	164	experimental	2.39	-0.55	x		x	1.50	x

Note. Pubtype=publication type; unpub=unpublished study; pub=published study; experimental=experimental design/randomized control trial/clustered randomized trial; quasi-experimental=group comparison design without randomization; one-group=non-experimental design with pre- and post-measure of an intervention group; ES social=effect size on improved social-emotional development; ES problem=effect size on reduced problem behavior; WS=workshop; <30h=professional development with 30 or less training hours.

Study variance was analyzed with CMA version 2 (Borenstein et al., 2005). A heterogeneity test showed some between-study variance ($Q = 204$; $df = 12$; $p < .000$; $T^2 = .054$; $SE = 0.038$). Across studies, 94 % of the observed variance was due to systematic variance ($I^2 = 94.12$). The results indicated the use of a random effect model and a meta-regression to explain study variance. Investigating the stability of the summary effect, the Orwin's Fail Safe N test points out that 8 studies with a mean effect size of zero are needed to decrease the summary effect size to a negligible impact of $g < 0.20$. Classic Fail Safe N analysis demonstrated that 1,405 studies with no results are needed to decrease the aggregated effect to $g = 0$. A test of sensitivity showed a robust summary effect that varied between $g = 0.37$ to $g = 0.52$, when one study was removed from the meta-analysis.

Analyzing publication bias, a chi-square test showed that the variance was not significantly related to the sample sizes of teachers ($\chi^2 = 935$; $df = 913$; $p = .299$), sample sizes of children ($\chi^2 = 1020$; $df = 996$; $p = .292$), publication type ($\chi^2 = 85$; $df = 83$; $p = .419$) or standard error ($\chi^2 = 6885$; $df = 6806$, $p = .248$). Within the data set, Egger regression (*Intercept* = 2.53; $SE = 1.79$; $p = .190$) maintained the chi-square results and showed out that there is no relationship between study estimates and size of study.

4.3.2. Meta-analysis and moderator analysis on social-emotional and behavioral skills

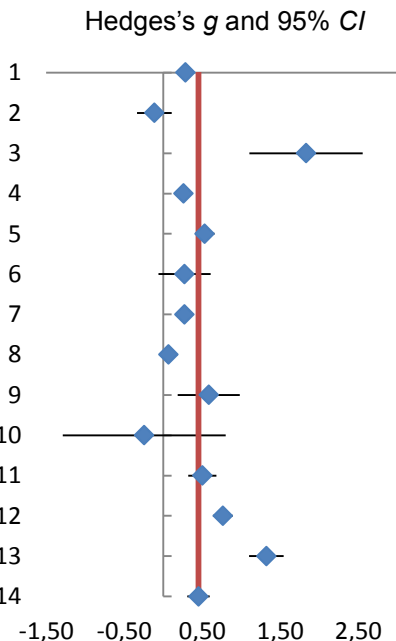
Using a variance-weighted random procedure, a small to medium summary effect of $g' = 0.45$ ($SE = 0.074$; $k = 86$; $CI = 0.31 - 0.60$; $Z = 6.131$; $p < .000$) could be found for social skills and reduced problem behavior. The confidence intervals for the summary effect range from small to medium intervention effects. The study variance showed that professional development programs were weighted in the meta-analysis, ranging from 1.61 to 9.99 % (see table 25). The aggregated treatment effect estimates show that 2 studies had negative impact estimates for professional development and 4 studies had lower confidence interval estimates in the negative range.

The aggregation of the findings from 13 studies including 86 effects revealed a significant summary effect of $g = 0.48$ ($SE = .130$; $k = 86$; $RIGLS = 251$; $\Delta\chi^2 = 639$; $p < .000$) for in-service professional development on positive social and behavioral development in young children. This is a small to medium effect according to Cohen's rules of thumb (1988). Distinguishing between social skills and challenging behavior, differential effects appeared related to subdomain (see table 26). A stronger professional development effect occurred for measures on social skills ($g = 0.55$) than for measures that showed a reduction in challeng-

ing behavior ($g = 0.33$). In particular, considerable in-service impacts were detected for peer interaction skills (e.g., social interaction and cooperation) with an effect size of $g = 0.71$, and for self-control abilities (e.g., motivation, persistence and attitudes) with an effect of $g = 0.66$. A significantly smaller effect of professional development ($g = 0.25$) was found for the reduction in externalizing problem behavior (e.g., distractibility, disruptive or aggressive behavior).

Table 25. Meta-analysis on social skills and problem behavior

Nr.	Study	ES	SE	Lower CI	Upper CI	Relative Weight	Hedges's g and 95% CI
1	Baker-Henningham et al.	0,28	0,04	0,21	0,36	9,82	1
2	Bryant et al.	-0,11	0,11	-0,34	0,11	8,18	2
3	Cordell	1,83	0,37	1,11	2,56	2,86	3
4	Domitrovich et al.	0,26	0,03	0,21	0,31	9,99	4
5	Gallagher et al.	0,53	0,06	0,41	0,66	9,41	5
6	Girard et al.	0,27	0,17	-0,06	0,61	6,56	6
7	Llyod & Millenky	0,28	0,03	0,22	0,33	9,98	7
8	Morris et al.	0,07	0,06	-0,04	0,18	9,54	8
9	Rhodes & Hennessy	0,58	0,20	0,19	0,98	5,74	9
10	Shernoff & Kratochwill	-0,24	0,53	-1,29	0,80	1,61	10
11	Snyder	0,50	0,09	0,32	0,68	8,74	11
12	Stoiber & Gettinger	0,76	0,06	0,64	0,89	9,40	12
13	Thanasetkorn	1,32	0,11	1,10	1,54	8,18	13
14	Summary effect	0,45	0,07	0,31	0,60		14



Through a random multi-level procedure, methodological and theoretical effect modifiers were investigated. First, all the variables were examined individually (model 1). Regarding publication bias, the multi-level moderator analysis showed that the aggregate summary effect was not moderated by publication type (*regression coefficient* = -0.167 ; $SE = 0.284$). However, the subgroup estimates suggested that the effects reported in journals ($g = 0.41$) were slightly smaller than those in unpublished papers ($g = 0.58$). Taking a closer look, the results from the doctoral dissertation by Cordell (2010) and Thanasetkorn (2009) clearly indicated a huge professional development effect ($g > 1.30$). Further, a file drawer problem can be rejected given the regression estimates for sample sizes of teachers (*regression coefficient* = -0.007 ; $SE = 0.004$) and children (*regression coefficient* = -0.001 ; $SE = 0.000$). Investigating methodological variables, comparison type and informant appeared as significant effect modifiers. The results on comparison type showed significant differences when effect sizes were obtained from gains made between pre- and posttest ($g = 0.74$). Unexpectedly, the reported effects were smaller when parents were asked to rate social or behavioral skills ($g = 0.32$) and higher when external observers rated these abilities ($g = 0.62$).

In model 2, the significance of theoretical effect modifiers was tested adjusting for informants (parents and observers). In model 3, the theoretical modifier was entered simultaneous with the informant variables and outcome domain (1=social). The methodological covariates and outcome domain correlated modestly with each other ($r < .22$). Model 2 and 3 analyze the independent influence of theoretical predictors, adjusting for methodological artefacts and outcome domain.

Table 26. Moderator analysis on social and behavioral outcome domains

Parameter	Social-emotional and behavioral abilities				
	<i>k</i>	<i>g</i>	RIGLS	Residual variance	$\Delta\chi^2$
Random effects					
Summary effect (k=86)		0.48*	251	.194	638*
Social development	44	0.55*	218	.188	525*
Social skills (composite)	16	0.50	251	.195	134*
Peer interaction	13	0.71*	205	.202	163*
Self-control	9	0.66*	247	.177	131*
On task behavior	8	0.43	250	.191	98*
Problem behavior	42	0.33*	218	.188	117*
Problems (composite)	8	0.44	251	.196	32*
Internal problems	11	0.42	259	.194	33*
External problems	21	0.25*	209	.191	126*

Note. * $p < .05$; RIGLS=restricted maximum likelihood estimation.

In addition to potential methodological covariates, *training duration and intensity* were examined as theoretical effect modifiers (see table 27). While some tendencies became obvious that training with shorter durations were more effective (*regression coefficient* = $-.077$; $SE = 0.035$), subgroup analysis showed that in-service formats with exactly 9 months' duration were significantly less effective than the other formats ($g = 0.21$). At first glance, regression estimates did not suggest that less intensive training is more effective (*regression coefficient* = $-.001$; $SE = 0.001$). However, after splitting the training amount into 2 categories with more or less than 30 hours¹⁸, estimates clearly indicated that shorter in-service programs ($g = 0.66$) are considerably more effective than long-term programs ($g = 0.21$).

Several *in-service delivery formats* have been used to foster social skills and reduce problem behavior at the same time. The comparison of results from the moderator analysis showed that it is not possible to clearly attribute effectiveness to a specific delivery type. Some tendencies emerged when comparing the combination of coursework and onsite support ($g = 1.00$) with all other formats ($g = 0.40$). However, the result (*regression coefficient* = $.605$; $SE = 0.377$) was not statistically significant.

¹⁸ A cut off was taken at 30 training hours for educational attainment that is equivalent to two credit hours at university or college course or 5 days of workshops with 6 hours a day.

Table 27. Moderator analysis of in-service effects on social and behavioral outcomes

Characteristic			Model 1		Model 2		Model 3	
	<i>k</i>	<i>g</i>	<i>k</i> =86		<i>k</i> =86		<i>k</i> =86	
			Residual variance	$\Delta\chi^2$	Significance of Moderator	Residual variance	Significance of moderator	Residual variance
Publication type				203*	n.s.	.145	n.s.	.173
Published	65	0.41	.22					
Unpublished	21	0.58*	.22					
Methodology								
Comparison (EG vs. CG)				572*	-	-	-	-
Yes	80	0.42	.20					
No	6	0.74*	.20					
Difference at pretest								
Experimental group higher	37	0.49	.20	141*	-	-	-	-
Control group higher	27	0.46	.19	80*	-	-	-	-
Informant								
Child (test)	4	0.42	.19	19	-	-	-	-
Teacher	50	0.49	.19	428*	-	-	-	-
Observer	21	0.62*	.23	124*	-	-	-	-
Parents	11	0.32*	.19	28*	-	-	-	-
In-service training intensity								
Duration								
Less than 3 months	12	0.78	.15	236*	n.s.	.195	n.s.	.188
3 to 6 months	29	0.54	.22	123*	n.s.	.248	n.s.	.243
9 months	45	0.21*	.14	239*	n.s.	.187	n.s.	.172
Amount				213*				
Less than 30 hours	57	0.66*	.14		n.s.	.160	sig.	.147
30 to 247 hours	27	0.21*	.14					
In-service delivery								
Course only				3	n.s.	.246	n.s.	.242
Yes	2	0.59	.21					
No	84	0.47*	.21					
Onsite support only				34*	n.s.	.228	n.s.	.226
Yes	8	0.14	.19					
No	78	0.56*	.19					
Workshop & onsite				393*	n.s.	.226	n.s.	.208
Yes	57	0.58	.21					
No	29	0.41*	.21					
Course & onsite				24*	n.s.	.223	n.s.	.220
Yes	5	1.00	.19					
No	81	0.40*	.19					
Workshop, course & onsite				143*	n.s.	.215	n.s.	.196
Yes	57	0.17	.19					
No	29	0.54*	.19					
Target child				372*	n.s.	.248	n.s.	.244
At risk	76	0.46	.22					
Not at risk	10	0.52*	.22					
Curriculum implementation								
Mayor focus of training	7	0.75	.22	0	n.s.	.244	n.s.	.245
Content part of the training	51	0.38	.20	171*	n.s.	.238	n.s.	.233

Note. * $p < .05$; n.s.=not significant; sig.=significant;

Model 1=unadjusted;

Model 2=adjusted for parents and observers as informants;

Model 3=adjusted for parents and observers as informants and outcome domain.

Children at risk were the focus of professional development that implemented behavioral management strategies or systems, social curricula or specific child-teacher-interactions in early childhood classrooms. The findings from a previously conducted systematic review re-

vealed that the evidence base for the effectiveness of professional development programs on children at risk is ambiguous. The findings from the meta-analytic investigation confirmed the lack of empirical clarity. So far, there is no evidence that at risk children ($g = 0.46$) benefit more than children not at risk when their teachers participate in professional development ($g = 0.52$). At this stage in research, compensatory effects caused by professional development on the development of children at risk cannot be assumed.

Exploratory analysis on the influence of the implementation of a specific or comprehensive curriculum on in-service training effect sizes on social and behavioral gains were carried out. A moderator analysis investigated whether *curriculum implementation* was a major target of the program (*regression coefficient* = $-.313$; $SE = 0.401$, $k = 86$) or curriculum/manual-based strategies were part of the training in combination with other training aspects (*regression coefficient* = -0.178 ; $SE = 0.268$; $k = 86$). However, the regression coefficients were not significant and did not indicate an influence of curriculum implementation on training effects. Only one study (Domitrovich et al., 2007) provided information on *fidelity*, to be more precise, on the extent to which the PATH curriculum was implemented in classrooms. Due to the lack of information, it was not possible to analyze the moderating influence of fidelity for the strategies, management techniques or curricula with meta-analytic techniques.

In model 2, none of the findings on training intensity or duration could be confirmed. Once again, after adjusting for informants, none of the professional development programs differed considerably in effectiveness. In model 3, outcome domain was additionally used as covariate. A low training intensity of less than 30 hours emerged as a significant effect modifier once more.

4.3.3. Narrative review of effective studies on social-emotional and behavioral skills

From the 13 studies, only 2 studies reported significant improvements in social behavior. The study by Cordell (2010), with an aggregated effect size of $g = 1.83$ based on change scores, focused on the development of English Language Learners in dual language sessions in Head Start classrooms. The technical assistant intervention for 3 teachers consisted of 2 major targets: 1) the implementation of the Creative Curriculum with high fidelity and 2) improvements in family literacy practice. The Creative Curriculum was implemented to foster children's social/emotional, physical, cognitive, and language development. The teacher intervention consisted of weekly conferences (60 minutes) for a period of 12 weeks and onsite support by a researcher who was available via telephone or at the center. During the 12 weeks, the teachers also received weekly lesson plans to introduce specific topics in the

classroom. Measured with the Creative Curriculum Developmental Continuum, the English Language Learners made significant gains from fall to spring in the targeted areas sense of self, responsibility for self and others, and prosocial behavior.

The study by Thanasetkorn, with an aggregated effect size of $g = 1.32$, evaluated the implementation of the “The 101’s: A Guide to Positive Discipline” curriculum to improve positive teacher-interactions and teacher-child relationship quality in Thai classrooms in Bangkok. Overall, 20 kindergarten teachers and 164 students from 2 private elementary schools participated. Six teachers participated in a half-day workshop plus 4 follow-up sessions and received the manual “The 101’s: A Guide to Positive Discipline”. The 6 teachers received the manual and the parents received a half-day training session on positive discipline. Eight teachers formed the control group in which teachers and parents did not receive any training. The training helped the teachers to increase their positive teacher-child interactions and to create an emotionally supportive environment. The students from the teacher training group had significantly higher scores in school liking, cooperative participation, and self-directedness than students from the control or parental training groups. At posttest, the students from the training condition had statistically lower values in school avoidance than the students in the control and parental training groups.

4.4. Link between in-service training, quality ratings and child development

In the integrative framework, it is assumed that in-service training affects students’ development through multiple steps (see section 2.1.1.). This section focuses on 2 of those steps. First, in-service training improves teaching and intervention practice as well as classroom quality ratings. Second, improved classroom quality affects student achievement. Earlier sections demonstrated that professional development improves the quality of child care (proximally) and fosters child development (distally). It was not possible to test quality improvements through professional development as an effect modifier on student gains in the primary meta-analyses, because only a limited amount of studies provided data on quality ratings and child development. This section focuses on the link between professional development, child care quality and child development. The main research objective of this section is to show how quality improvements that appear during in-service training are related to in-service training effects on child development. An exploratory analysis was conducted to investigate if quality improvements predict in-service training effects on child development. To analyze this link, studies were needed that measure both quality improvements and student gains through professional development.

4.4.1. Description of primary studies on quality ratings and child development

Overall, 9 studies including 10 treatments evaluated the impact of in-service programs on quality ratings and child outcomes simultaneously. Nine studies provided data on 68 contrasts to estimate effect sizes. Of these, 53 effects were based on language and literacy scores, 8 on social-behavioral ratings, 6 on assessment of cognition, knowledge and school readiness and one on early math testing. All the studies were conducted in the United States and were published recently (2006 - 2011). Overall, data from 486 teachers and 4,504 children were included in the meta-analysis. Variations existed in the size of teacher samples, varying between 18 and 108 participants, and for child samples with a minimum of 193 and a maximum of 808 students at pretest (see table 28). Regarding the number of students, all studies evaluated in-service training effectiveness on a large scale. To measure the impact on quality ratings, the ELLCO was mainly used ($n = 7$). Additionally, quality was measured through the Environmental Rating Scales ($n = 2$), the CLASS at classroom level ($n = 3$) and the inCLASS at target child level ($n = 1$). The 9 in-service programs consisted of intensive professional development approaches with durations of 4 to 24 months and 30 to 216 training hours. Most of the treatments ($n = 7$) used a combination of delivery methods (e.g., workshop, course and onsite support). All the treatments, except the workshop approach by Lonigan et al. (2011), applied onsite support to improve quality and foster child development simultaneously.

4.4.2. Meta-analyses on in-service training effects on quality and child development

Meta-analysis of in-service training effects on quality ratings

For the aggregation of in-service effects on quality ratings, 82 effect sizes were available (range $g = -0.39$ to 2.10). The aggregation of effect sizes per treatment showed that treatment effects varied according to quality ratings ranging from $g = -0.39$ to 1.71 . Only the study by Clancy-Menchetti (2006) reported a negative effect of in-service training on quality ratings. The meta-analysis revealed a significant weighted summary effect of $g' = 0.45$ ($SE = 0.108$; $k = 82$; $CI = 0.24$ to 0.67 ; $Z = 4.208$; $p < .000$). To estimate the summary effect, studies were weighted in relation to their variance, ranging from 6.13 to 12.51 % (see table 29).

The heterogeneity test showed some between-study variance ($Q = 82$; $df = 9$; $p < .000$; $T^2 = .092$; $SE = 0.069$). The total variability in effect sizes between studies is 89 % ($I^2 = 89.01$). Classic Fail Safe N estimates show that 289 studies with no effects are needed to decrease the aggregated effect to $g = 0$. With a range of $g = 0.34$ to 0.51 , the summary effect on quality ratings was quite robust when one treatment was removed.

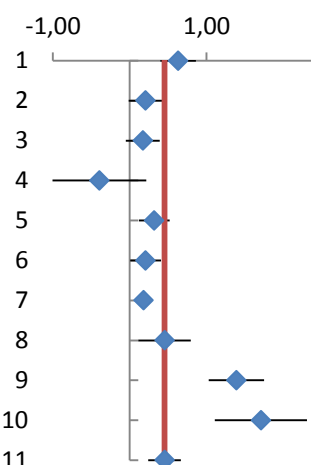
Table 28. Overview of studies on the impact on quality ratings and child outcomes

Author	Year	Country	Pubtype	Design	Sample teacher	Sample children	Scale	WS	CS	COP	ON	Duration (months)	Amount (hours)
Barnett et al.	2008	USA	pub	CRT	18	218	ERS, CLASS	x		x	x	9.00	59.00
Bryant et al.	2009	USA	unpub	RCT	108	236	ERS				x	9.00	30.00
Buyse et al.	2010	USA	pub	CRT	55	193	ELLCO	x		x	x	9.00	40.00
Clancy-Menchetti	2006	USA	unpub	CRT	40	500	ELLCO	x	x		x	24.00	216.00
Lonigan et al. (a)	2011	USA	pub	CRT	51	808	ELLCO	x			x	9.00	68.00
Lonigan et al. (b)	2011	USA	pub	CRT	51	808	ELLCO	x				9.00	32.00
Morris et al.	2010	USA	unpub	QNE	52	623	(in)CLASS	x	x		x	9.00	162.00
Onchwari & Keengwe	2010	USA	pub	QNE	44	626	ELLCO				x	5.00	40.00
Powell et al.	2010	USA	pub	RCT	88	759	ELLCO	x			x	4.00	37.00
Wasik & Hindman	2011	USA	pub	CRT	30	541	ELLCO, CLASS	x	x		x	9.00	43.75

Note. Pubtype=publication type; pub=published in journals or as book chapter; unpub=unpublished paper e.g., technical paper, dissertation or governmental report; design=study design; RCT=randomized control trial assigning teacher to groups; CRT=clustered randomized trial assigning classrooms or center to groups; QNE=quasi-experimental design without randomization or matching/non-equivalent groups design; Scale=quality rating scale; WS=workshop; CS=course and in-service with ongoing sessions; COP=community of learning or community of practice; ON=onsite support like coaching, consulting, or mentoring.

Table 29. In-service training impacts on quality ratings

Nr.	Study	ES	SE	Lower CI	Upper CI	Relative weight
1	Barnett et al.	0.63	0.12	0.40	0.86	11.07
2	Bryant et al.	0.21	0.11	-0.01	0.42	11.22
3	Buysse, et al	0.17	0.19	-0.01	0.42	11.22
4	Clancy-Menchetti	-0.39	0.31	-1.01	0.22	6.13
5	Lonigan et al. (a)	0.32	0.10	0.12	0.52	11.40
6	Lonigan et al. (b)	0.21	0.10	0.00	0.41	11.36
7	Morris et al.	0.18	0.04	0.11	0.25	12.51
8	Onchwari & Keengwe	0.45	0.17	0.11	0.80	9.54
9	Powell et al. (a)	1.39	0.19	1.03	1.75	9.25
10	Wasik & Hindman	1.71	0.31	1.11	2.31	6.29
11	Summary effect	0.45	0.11	0.23	0.66	

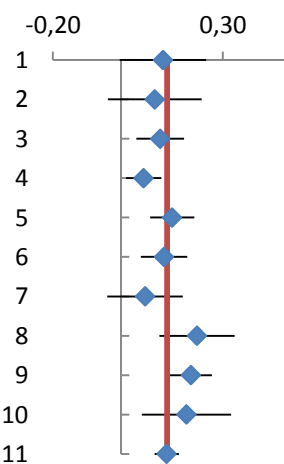
Hedges's g and 95% CI**Meta-analysis of in-service training effects on child outcomes**

Overall, 68 effects on different child outcomes could be estimated; these ranged from $g = -0.24$ to 0.44 . The aggregated professional development effects fell in a very moderate span, ranging from $g = 0.07$ to 0.22 . The weighted summary effect for the 9 studies is in the lower range ($g' = 0.14$; $SE = 0.019$; $k = 68$; $CI = 0.10$ to 0.17 ; $Z = 7.304$; $p < .000$). Studies were weighted according to study variance, ranging from 5.38 to 15.77 % (see table 30).

Investigating study variance, a heterogeneity test showed some between-study variance ($Q = 17$; $df = 9$; $p = .054$; $T^2 = .001$; $SE = 0.002$). Overall, 46 % of the observed variance was due to systematic variance ($I^2 = 46.06$). Classic Fail Safe N analysis demonstrated that 252 studies with no results were needed to decrease the aggregated effect to $g = 0$. The summary effect on child outcomes was very robust, ranging from $g = 0.12$ to $g = 0.14$ when one treatment was removed from the meta-analysis.

Table 30. In-service training impacts on child development

Nr.	Study	ES	SE	Lower CI	Upper CI	Relative weight
1	Barnett et al.	0.12	0.07	0.00	0.25	6.05
2	Bryant et al.	0.10	0.07	-0.04	0.24	5.38
3	Buysse. et al	0.12	0.04	0.05	0.19	12.46
4	Clancy-Menchetti	0.07	0.03	0.01	0.12	15.77
5	Lonigan et al. (a)	0.15	0.03	0.09	0.22	13.30
6	Lonigan et al. (b)	0.13	0.04	0.06	0.20	12.73
7	Morris et al.	0.07	0.06	-0.04	0.18	7.32
8	Onchwari & Keengwe	0.22	0.06	0.11	0.34	7.35
9	Powell et al. (a)	0.21	0.03	0.14	0.27	13.85
10	Wasik & Hindman	0.19	0.07	0.06	0.32	5.73
11	Summary effect	0.14	0.02	0.10	0.17	

Hedges's g and 95% CI

4.4.3. Association between quality improvements and in-service training effects on child development

The results from the primary studies and meta-analyses demonstrated that it is possible to increase quality scores and child development through in-service professional development simultaneously. As in the previous results section, the (proximal) summary impact of in-service training on quality ratings was larger than the (distal) summary impact on child development. Within the integrated framework and other path models (cf. Fukkink & Lont, 2007; Yoon et al., 2007), multiple steps of change and improvement are needed to attain increased student development. First, professional development improves teachers' knowledge, attitudes and skills. Second, better knowledge and skills improve teaching and classroom quality and third, improved teaching fosters child development. Within the scope of this dissertation, the findings on step one – the mechanism which facilitates teacher learning that results in better knowledge and skills – cannot be identified. Nevertheless, the meta-analysis of in-service training on quality ratings provided evidence for systematic changes in classroom practices at step 2. For step 3, a linear regression was conducted to analyze how quality improvements that occur during training predict in-service training effects on child development (see table 31).

Table 31. Regression analysis on quality improvements predicting child development

	<i>B</i>	<i>SE B</i>	β	p
In-service training effects on quality ratings	.06	.02	.73	.017
R^2	.532			

In the regression model, the aggregated effect sizes per treatment on quality ratings and the aggregated effect sizes on child outcomes were used. Quality gains during training predicted the in-service training effects on child outcomes ($\beta = .73$; $p = .017$). Fifty-three % of the variance in the effect sizes on child development were explained by in-service training effects on quality ($R^2 = .53$). The model explains a substantial proportion of student achievement gains, whereas the summary effect on child development is meager. This indicates that tremendous professional development efforts are needed to achieve significant gains in child development.

4.5. Evaluation of in-service training components summarized in the integrative framework

The integrative framework includes effect modifiers that have been discussed in the literature as influencing in-service training effects on quality ratings and child development. This chapter focuses on the evaluation of the components summarized in the integrated framework. According to Baumeister and Leary (1997), conclusions can be made with different levels of confidence through systematic reviews and meta-analyses. Through moderator analysis, the null hypothesis is tested and (dichotomous) conclusions can be drawn (supporting or rejecting the null hypothesis). Based on previous meta-regressions and subgroup analyses (see sections 4.1.2.; 4.2.2.; 4.3.2.), conclusions are based on current evidence. They confirm that a hypothesis is correct when modifiers become significant (see summary in section 4.5.1.)

Narrative reviews of the most effective in-service programs were conducted to provide additional information on best practice models and effective components (see sections 4.1.3.; 4.2.3.; 4.3.3.). These best practice models and their components can be adduced as restricted evidence. The components retrieved from narrative reviews are potential effect modifiers indicating that the hypothesis was not conclusively confirmed, but is currently the best guess (cf. Baumeister & Leary, 1997).

In addition, several professional development components that have been discussed in the literature did not emerge as significant effect modifiers in moderator analyses or narrative analysis. The hypotheses cannot be rejected or supported, because the current evidence permits no conclusions for several reasons. Findings on these components were either internally inconsistent or there was a lack of evidence.

To sum up, only very few components summarized in the integrative framework could be statistically confirmed as significant effect modifiers (see figure 10). The narrative review, based on primary studies, also suggests some variables from the integrative framework as potential effect modifiers. For some variables, inconsistent findings were found in the moderator analyses and it is not possible to draw clear conclusions. However, most of the components of professional development have not yet been systematically and statistically tested. There is an urgent need for more research studies testing effect modifiers mentioned in the literature. In particular, training transfer variables known from adult education research have not been included in the description of training studies in the early childhood field.

True	WHO Teacher working with children at risk	HOW Training intensity Training duration	WHAT -
Best guess	WHO Trainer characteristic - Pretraining	HOW Delivery - Formats with guidance - Formats with feedback Principles of Learning - Performance based - Active learning	WHAT Core competencies - Content/domain specific
Inconsistent	WHO -	HOW Delivery - Scale-based approaches - Mixture of delivering formats - Coaching	WHAT Curriculum
Lack of evidence	WHO Trainer characteristic - Experience - Education Trainee characteristics - Education - Motivation - Abilities - Personality - Language/ethnicity	HOW Delivery - Sequencing - Fidelity/quality of training Principles of learning - Content relevance - Difficulty level - Defining learning goals - Collective participation	WHAT Core competencies - Defined by early learning standards and credentials - Coherent with curriculum & learning standards

Figure 10. Evidence-based conclusions on in-service training effect modifiers

4.5.1. Significant training effect modifiers of professional development

Several components and features of professional development have been assumed in literature as modifying in-service training effects (see section 2.2.1 and 2.2.2). Based on the findings from the meta-regressions and subgroup analyses, the moderating role of in-service training components was examined. The results from the 3 meta-analyses are summarized in table 32.

Methodological variables as effect modifiers

Several aspects of research design features (e.g., sample size, type of comparison group, randomization, contrast vs. pre-post, and methodological quality), respondent characteristics (e.g., age, gender, ethnicity, ses, ability group), and measurement features (source of information, researcher-developed measure, construct, and operationalization) have been discussed in the literature as modifying treatment effects (cf. Wilson & Lipsey, 2001).

Emphasizing *research design features*, type of contrast was a significant modifier within both meta-analyses on student development. Tendencies became evident in which effect sizes

obtained from posttest contrasts between EG and CG were significantly smaller and effect sizes made of pre-post gains were considerably larger. When descriptive data show differences in mean values at pretest between EG and CG, the in-service training effect was significantly related to the group with higher values, at least for the meta-analyses on quality ratings and language and literacy skills. Random assignment was not a significant moderator. Studies with or without random assignment of teachers to training or control conditions had similar results. Dropout rate was mostly described vaguely or could only be coded and calculated from tables that reported sample size for pre- and posttest. Therefore, a cut off of 25 % was used, but significant differences due to attrition rate did not emerge. At effect size level, good reliability (measured through Cronbach's $\alpha > .70$) and high inter-rater agreements (reported either through *Kappa* or *ICC* $> .80$) were dichotomously coded, because standardized effect sizes tend to be larger when the measurement error is smaller and measures are reliable (Burchinal, 2008). In the meta-analysis on quality ratings, effect sizes with good reliability values were significantly larger than effect sizes without.

After the *operationalization of dependent variable*, several effect modifiers emerged in the 3 meta-analyses. Different external rating scales reflect different *constructs* of early childhood quality. In particular, the more interaction-focused constructs used by the CLASS resulted in significantly larger effect sizes. Several instruments and scales were used to measure in-service training effects on language, literacy, and social or behavioral development. Some studies reported only at subdomain or subcategory level and some reported total scores. Effect sizes on language and literacy reported in *total scores* were smaller than in subtests. Several meta-analyses have demonstrated that source of information or informant explained a large proportion of the effect size (cf. Blume et al., 2010; Wilson & Lipsey, 2001). A notable influence by informants was found only in the meta-analysis of in-service training impacts on social and behavioral development; here, observer ratings resulted in larger, and parental ratings in significantly smaller effect values.

Outcome domain as effect modifier

In addition to the operationalization of dependent variables, significant differences in training effectiveness were found for several outcome domains. These significant differences may have appeared due to the constructs or the sensitivity of the scales to measure changes. Further, variations in effect sizes might have occurred due to aspects of trainability (effort, duration etc.). In the areas of language and literacy, in-service programs had larger impact on literacy-related skills than on language-related abilities. In particular, students benefited more in the developmental domains of elision, concepts about print, and writing when teachers received in-service training. Further, the effect on language comprehension abilities was significantly larger than on other language domains. The impact of professional development

on phonological awareness was tested only by one study, but the effect size was significantly smaller than the in-service training impact on other developmental domains. In addition, the results showed that it is difficult to improve non-word repetition and receptive vocabulary through professional development. In the area of social-behavioral aspects of development, in-service training had larger effect sizes fostering social-emotional abilities than reducing problem behavior. In particular, in-service training was more effective in improving peer interactions and self-control abilities, while only modest effects were found for external problem behavior.

Table 32. Summary of findings on potential effect modifiers

Meta-analysis	Methodology	Outcome domain	Context	Training
Quality ratings	Renowned scale + EG higher at pretest + CG higher at pretest – Cronbach's $\alpha > .70$ + CLASS +			45 - 60 h of training + Only coaching +
Language & Literacy	Posttest contrast – EG higher at pretest + CG higher at pretest – Total score -	Language + Receptive vocabulary - Comprehension + Non-word repetition – Literacy + Elision + Phonol. awareness – Concepts about print + Writing +	At risk +	15 – 30 h of training + Workshop & Course + Mixed delivery +
Social skills & problem Behavior	Informant: observer + Informant: parents -	Social skills + Peer interaction + Self-control + Problem behavior – External problems +		Training = 9 months + <30 hours of training +

Note. +=positive significant moderator; -=negative significant moderator.

Contextual features as potential in-service training effect modifiers

Teachers participating in professional development are diverse in their backgrounds (e.g., education, experience, and role) and work in diverse settings regarding the grouping of children (age and at risk factors). For language and literacy outcomes, moderator analysis showed that in-service training effects were larger for teachers working in child care settings, where classrooms consisted mainly of children from disadvantaged families. Other contextual factors did not emerge as modifiers. This finding suggests that conclusions on professional development effectiveness can be made, regardless of setting type and context.

Training intensity as an effect modifier

Based on subgroup analyses and meta-regressions, conclusions on training effects due to differences in training intensity and delivery format can be drawn. All meta-analyses contained short-term and long-term programs with a training amount of up to 200 training hours. Findings showed that in-service programs with 45 to 60 training hours were the most effective in changing classroom quality. To translate professional development efforts into gains in student achievement, even less training hours were needed. In-service programs with less than 30 hours were very effective in fostering social abilities and reducing challenging behavior and, in particular, programs with 15 to 30 hours resulted in strong teacher training effects on language and literacy abilities.

4.5.2. Potential effect modifiers from the narrative review

Not all the variables from the integrative framework could be conclusively confirmed through statistical meta-analytic procedures. The narrative reviews of the highly effective professional development programs provide some insights on potential effect modifiers. The following conclusions are based on a certain number of primary studies that reported large professional development effects.

This applies to the *training of in-service training providers* prior to the intervention. Narrative analysis on the most effective in-service programs suggested that trainers were highly skilled and knowledgeable of the content they provided. Further, the trainers received a certificate or were the authors/developers of the in-service program.

Due to insufficient reporting, it was not possible to code the specific content of professional development with high inter-rater reliability. The findings from the narrative analysis showed that the most effective programs were *(domain-)specific and content-focused*. Within the programs, the teachers learned to implement new interaction strategies and activities in classrooms.

The findings from the narrative analysis on the most effective programs indicated that several *didactic principles* are the adjusting screw to achieving larger in-service training impacts. In particular, approaches with *performance-based assignments* that combine direct skill training with field experience seem very promising. Further, the programs demonstrated the target strategy through *videos* or *modeling*. The in-service training provided direct, distal or video *feedback* and offered *active learning opportunities*.

4.5.3. Inconclusive effect modifiers and lack of evidence

Several other components of professional development are listed as in-service effect modifiers in the literature, but the current evidence permits no conclusions. Whether these hypotheses and assumptions should be supported or rejected cannot be made due to 4 reasons.

First, conclusions cannot be made with confidence, because the *current state of the evidence is inconsistent* and moderator analysis has not come up with clear results. Referring to the results from the subgroup analysis on delivery type, various and inconsistent findings on the benefits of specific delivery formats emerged. The findings showed that delivery format influences how well the professional development works, but different delivery formats are needed to obtain different professional development effects on a variety of outcomes. For quality improvements, individualized coaching approaches seemed to be the most effective strategies for enhancing classroom practice. The results from the moderator analysis on language and literacy outcomes showed different findings. In particular, the combination of workshop and course had significantly larger effects on language and literacy scores; but this finding emerged due to one outstanding impact from the study by McGill-Franzen et al. (1999) that used this format solely. Excluding the study by McGill-Franzen et al. (1999), the combination of coursework and onsite support seemed to be the most promising in-service training approach. Nevertheless, with or without the results from McGill-Franzen et al. (1999), approaches that mixed different formats with onsite support were negatively associated with teacher training impact on language and literacy abilities. In the meta-analyses on social skills and challenging behavior, no significant modifiers concerning delivery format were found. Based on the current state of research, the hypotheses that a specific delivery format (e.g., customized or manual-based) or approach (either modern, traditional or a combination of both) is significantly more effective for all outcome levels than the other approaches must be rejected. Instead, specific approaches were effective for specific outcomes.

Scale-based approaches are they are considered as training to the test and treated as the foundation for in-service training goals and feedback. The assumption that in-service programs that use scale-based approaches are more effective than others are yielded particularly inconsistent results. In addition, moderator analyses on the implementation of a curriculum or specific activities in the classroom as a major target of professional development did not provide evidence that this modified training impacts on quality improvements or student gains. The coefficients for any kind of curriculum implementation were negative, but not significant.

Second, *studies did not report sufficient information* or studies provided ambiguous information so that it was not possible to code them reliably and use these modifiers in meta-regressions. Several challenges occurred regarding the “what-component” in trying to code

the exact content of professional development. Notably, the core competencies and learning outcomes/goals for teachers or target behaviors in in-service programs were not clearly stated in most of the studies. In line with the exact content, it was not reported whether teachers received education credentials, an attestation of qualifications or competences for attending the training or not. Having ambiguous or undefined goals, contents, and principles of professional intervention, it is hard to operationalize and investigate professional development quality or at least fidelity. Not surprisingly, information on quality and fidelity of training was lacking in most of the studies. Additionally, it was not possible to code rhythm of intervention (e.g., weekly, monthly, or by session) with high reliability due to a lack of information. While it is widely presumed that principles of learning (e.g., content relevance, difficulty level, defined learning goals, and collective participation) modify professional development effects, they were not clearly stated in the documents. Contextual variables such as the underlying curriculum or center characteristics were insufficiently described by the authors. Hence, it is not possible to analyze whether the professional development was coherent with respect to the curriculum, learning standards or center characteristics or not.

Third, there is a *general lack of evidence*, because studies did not gather information on training transfer factors. This particularly concerns the “who-component” of the in-service training. Studies did not report data on trainee and trainer characteristics. Information on training motivation, self-efficacy, or the experience of trainees or trainers was not reported. Further, there was a lack of facts on aspects to do with the grouping of participants (e.g., starting level, degree, team approach) and relational issues of matching between trainees and trainers.

Fourth, some *detailed categories were missing in the coding schema*. Unfortunately, only the aspect of material provision was included in the coding scheme, but not the aspect of financial support and compensation. Findings from the narrative analysis on training impacts on classroom quality pointed out that some of the highly effective programs received financial support and/or classroom material. This financial support could be given as quality improvement grants, funds for professional development (e.g., stipends or tuition costs), monetary support for expenses for release time and financial funds for curriculum materials and classroom equipment (Brinks, 2007; Boller et al., 2010; Grace et al., 2008; Palsha & Wesley, 1998). This similarity brings up the question whether financial support, as another component of professional development, has a potential influence on training impacts on quality ratings.

5. Discussion

5.1. Discussion of results

1. Professional development effectiveness

The purpose of this investigation was to examine the effects of professional development for early childhood educators on external quality ratings and child outcomes. In particular, 2 aspects of teacher training effects were analyzed. First, the magnitude of the aggregated effect of in-service training programs was estimated for quality ratings ($g' = 0.68$; CI = 0.55 to 0.82), language and literacy scores ($g' = 0.38$; CI = 0.29 to 0.47), and social, emotional, and behavioral development ($g' = 0.48$; CI=0.31 to 0.59). The findings demonstrate that in-service professional development matter.

a) Magnitude of in-service training effects

The magnitude of small to medium *effect size* is in line with other meta-analyses and systematic reviews that have synthesized the findings on professional development impacts systematically. The review by Isner et al. (2011) analyzed the effects of coaching models. A large proportion of studies ($n = 27$) reported some positive coaching effects on external quality ratings (e.g., ERS, ELLCO, or CLASS). In the review, 21 studies provided information on student achievement. Of them, 12 studies reported positive impacts on language and literacy and 6 suggested coaching effects on behavioral measures. Based on a limited amount of studies published between 1980 and 2005, the meta-analysis from by Fukkink and Lont (2007) shows positive training effects at caregiver level ($ES = 0.45$; $k = 78$) and at child level ($ES = 0.55$; $k = 15$). The review by Yoon et al. (2007) of 9 studies published between 1986 and 2003 suggests a positive impact of in-service training for K-12 teachers on student achievement in kindergarten and elementary school. Considering the effectiveness of in-service programs in an isolated way, great *variability and heterogeneity* were found in the results: some studies demonstrated no effects, positive or negative effects. In their review, Zaslow and colleagues (2010a; 2010b) identified several professional development programs that targeted the promotion of language and literacy ($n = 22$), social and behavioral development ($n = 14$), and classroom practice and quality ($n = 10$). In line with the findings from the meta-analyses, evaluation studies reported mixed effects for educator practice and for child outcomes.

b) Magnitude of estimated professional development impacts on students' learning in comparison with curriculum effects and comprehensive early education programs.

Comparing in-service training effects with the impact of preschool curriculum on quality ratings and student achievement, the Preschool Curriculum Evaluation Research Initiative (2008) provides the most comprehensive overview of what really works. Within the PCER-Study, 14 curricula were implemented in 15 sites and evaluated with standardized measures of quality (CIS; ECERS; TBRs) and child development in the areas of language, phonological awareness, reading, social behavior and mathematics. The report demonstrated that the findings from end-of-pre-kindergarten and kindergarten were widely diverse, depending on curriculum. Curriculum impact on quality ratings was investigated. The implementation of different curricula showed that ECERS-R scores were positively and negatively affected (distribution of $ES = -0.71$ to 1.66). Ten evaluations indicated a notable curriculum effect of $ES > 0.25$ on quality ratings. Curriculum implementation results are inconclusive for child outcomes, showing positive and negative effects on receptive vocabulary ($ES = -0.38$ to 0.48 in the PPVT), language abilities ($ES = -0.22$ to 0.46 in the TOLD), phonological awareness ($ES = -0.17$ to 0.38 in the Pre-CTOPPP), social skills, ($ES = -0.44$ to 0.42 in the SSRS), problem behavior ($ES = -0.31$ to 0.50 in the SSRS) and learning behavior ($ES = -0.44$ to 0.30 in the PLBS). In the results of 15 curriculum evaluations, only a limited number of studies reached an appreciable value of $ES > 0.25$ on student achievement.

The aggregated in-service effects estimated within the current meta-analyses on quality ratings and student achievement are above the cut-off for a significant impact ($ES > 0.25$). The summary effects on language and literacy ($g = 0.40$) as well as the aggregated effect size on social and behavioral development ($g = 0.48$) were at the high end of the estimated curriculum effects. One potential explanation of why in-service training effects are larger is that through in-service professional development, the quality of the learning experience during daily routines is increased. The summary in-service effect on quality ratings showed that learning experiences in classrooms are notably improved. Further, the findings from the fourth meta-analysis demonstrated that student gains were predicted by changes in classroom practice to a certain extent. The findings suggest that changing student-teacher interactions, teaching, and interventions require more than just providing a curriculum that includes evidence-based instructions and practice and requiring teachers to use it. Findings show that in-service programs were reasonably effective when the quality of daily routines changed. This includes the extensive use of language and interaction-promoting strategies, child-center strategies and the implementation of behavior management systems. The use of teacher-student interaction methods might improve the general amount of support and high-quality learning experience during circle, center, and free play time. Therefore, the amount and time of learning experiences through strategies that are integrated in daily routines

should be higher than the amount of learning experiences in programmatic curriculum lessons.

In addition to curriculum effects, an exploratory comparison of the impact of professional development and comprehensive education programs can be made. Comprehensive programs have been developed and implemented for disadvantaged children and families and combine, in the majority of cases, interventions for parents and early childhood educators (e.g., Head Start or Perry Preschool). The effect size from the current meta-analysis with a small impact ($g' = 0.45$) on social-emotional development and reduced problem behavior is slightly larger than the aggregated impact of comprehensive programs (see Anderson et al. 2003 [$ES = 0.38$]; Blok et al., 2005 [$ES = 0.02$]; Camilli et al., 2010 [$ES = 0.16$]; Nelson, Westhues, & MacLeod, 2003 [$ES = 0.27$]). This finding indicates encouraging evidence that in-service training is an option to improve classroom behavior overall and reduce challenging behavior. Meta-analyses on the impact of these programs show small to moderate positive program effects on cognitive development, for example, IQ, achievement tests, and ratings of academic skills (see Anderson et al. 2003 [$ES = 0.35$ to 0.43]; Blok et al., 2005 [$ES = 0.32$]; Camilli et al., 2010 [$ES = 0.23$]; Nelson, Westhues, & MacLeod, 2003 [$ES = 0.30$ to 0.52]). However, findings on program effects on language and literacy development were not reported, though the magnitude of in-service training impacts on language and literacy are similar to the program effects on cognitive or academic achievement scores.

2. The link between in-service training, quality improvement and child development

The findings from the meta-analysis clearly indicate the link from in-service training to child development through changes in classroom practice and quality. In particular, improvements in classroom quality through professional development explained more than 50 % of changes in student achievement. This indicates that in-service training effects on student development cannot appear in isolation from significant changes in classroom practice. It requires tremendous changes in classroom practices ($g' = 0.45$) to gain some benefits for student learning ($g' = 0.14$). This finding underpins path models that emphasize a two-step linkage: the link from in-service training to classroom practice and the link from improved practice to child development (cf. Klein & Gomby, 2008; Yoon et al., 2007). These findings substantiate the demand for professional development with a special focus on teacher practice. This initial conclusion can serve as a starting point to understanding the mechanism of professional development effects. However, the analysis was limited to 9 studies providing data on both changes in classroom quality and child development. Seeking to understand how in-service training effects work, teacher learning and changes in knowledge and skills need to be investigated simultaneously with classroom practice and student outcomes (cf. Fukkink & Lont, 2007; Hamre et al., 2012; Harvard Family Research Project, 2006). In accordance with the

conclusion by Zaslow et al. (2010b), we are at an early stage of professional development research. Further investigations are needed that report the different outcome levels: teacher knowledge and skills, classroom quality, and child development.

3. *Methodological effect modifiers*

A fundamental task of meta-analysis is the investigation of factors in study design, training design or context that moderate effectiveness. The moderator analyses came up with several *methodological effect modifiers*. In line with other research, pretest differences, reliability of scales, the aggregation level of outcomes (total or subscale) and different informants moderated the effect size significantly (see Beelmann, 2014; Wilson & Lipsey, 2001).

A practical implication for training design is provided by the finding that the magnitude of in-service training effects differed by the *outcome domains* that were tested. For example, significantly larger effects were found for professional development programs that targeted language and literacy abilities in the area of language comprehension, elision, concepts about print and early writing. Smaller effects appeared for receptive vocabulary, non-word repetition and phonological awareness. Domain-specific findings were found within the meta-analyses on social development and challenging behavior. Significantly larger in-service training effects were found for peer interactions, self-control and the reduction of external problem behavior. The differences in effect size may have occurred due to the following reasons.

First, the strategies and curricula implemented through in-service training did not target and foster all the developmental domains. The content of training was in line only with some outcome variables. Professional development that was more closely aligned with the measures produced larger effect sizes and less alignment of treatment and measures produced smaller effects (see Fukkink & Lont, 2007).

Second, the complexity of the abilities that were measured through different tests and assessments varied widely. For example, phonological awareness measured with the Phonological Awareness Task (PAT) includes rhyming, segmentation, blending, decoding, isolation of sounds, substitution of phonemes, and sound correspondence. As an example, speech production can either be assessed by the single task of producing single words in relation to a picture (EOWPT) or the production of full sentences in relation to a picture (SETK 3-5 ESR-Test). Changes in behavior and competencies are easier to measure, when they are less complex.

Third, some outcome domains seem to be more sensitive to the environment and training than others are. Oral language, such as expressive and receptive vocabulary, is known to correlate with exposure to reading or family literacy activities (Mol & Bus, 2011). Memory tasks such as non-word repetition are relatively impervious to training (Gathercole, 2006). The trainability of outcome domains might explain differences, too.

4. Training intensity is matched to the content of professional development

As mentioned earlier, *professional development amounts and duration* vary widely. Moderator analyses clearly demonstrated that the extent of in-service training matters, but more intensive training is not better. Conveying a specific content requires adequate intensity and timelines. The findings suggest that the amount and duration of in-service training is matched to its content and objectives. To attain significant improvements in classroom quality more time is needed. In particular, programs with a duration of 45 to 60 hours showed reasonable changes in classroom practice. In line with the findings by Zaslow et al., (2010a; 2010b), programs with less than 30 hours of training result in positive child outcomes. This indicates that short-term programs are sufficient when a set of specific skills is targeted. Hamre and Hatfield (2012) suggest that long-term and intensive training is needed when the professional development focus is broad and comprehensive.

5. Didactic principles matter

The findings from the narrative analysis on the core features of professional development are in line with policies on competence or performance-based guidelines (Deutsches Jugendinstitut & Weiterbildungsinitiative Frühpädagogische Fachkräfte, 2014; U.S. Department of Education, 2006). To be highly effective, professional development for teachers needs to move from static, knowledge-based workshops to active learning formats. These active formats focus on support, guidance and feedback for teachers to improve their daily interactions with children. Individualized modeling and feedback can be provided through onsite support, distally, through the internet, or by incorporating new techniques in workshops or coursework (cf. Zaslow et al. 2010a; 2010b). In summary, the body of research on in-service training effects is at an early stage and not all practice-focused approaches have shown positive impacts. The identification of effective features of professional development needs more advanced studies with clear definitions of the in-service models applied in practice.

5.2. Limitations

Several challenges and boundaries appear when conducting meta-analyses and systematic reviews. Research syntheses should extend knowledge on a specific research phenomenon by combining and comparing primary studies. As with primary investigations, research synthesis has to face several challenges in measuring causality, estimating the degree of generalization of findings and interpreting effect sizes. Further, meta-analyses are completely dependent on primary studies and their reports.

1. *Measuring causality and effects*

The main objective of the meta-analysis of experimental studies is to *verify cause and effect*, and in this case, the impact of in-service professional development programs on quality ratings and child outcomes. The internal validity of primary research, ascertained, for example; through randomization procedures, provides confidence that the independent variable actually affected the dependent variable. However, when most of the primary studies have low validity (e.g., non-random assignment, implementation of treatments on pre-existing groups) questions on causation cannot be confidently addressed. The general assumption is that RCTs provide evidence on the relationships between treatments and outcomes when several factors are controlled. Non-randomized evaluations offer hints on the relationship between interventions and outcomes, but are not controlled for several factors. Non-randomized studies reflect the way an intervention was applied on a day-to-day basis (Borenstein et al., 2009), but show causal connections less confidently. In the coding schema, information on randomized allocation of teachers and children to experimental or control conditions was collected in order to carry out a moderator analysis to assess the level of confidence on causality. The moderator analyses showed that the effects were not modified by the assignment procedures and randomization.

2. *Generalization of findings*

When *findings* are generalized, this refers to the transferability of results from a specific study using particular subjects, situations and procedures to a larger set of studies or to the population in general. The conditions in which the study has been conducted give a sense of the *external validity* of the results and effects reported in it (Hall et al., 1994). A meta-analysis can directly address the issue of generalization in 2 ways. Either use high standards for eligibility criteria (RCT with randomization, low attrition and group equivalence published within relevant time ranges) to include studies in meta-analyses (see WWC, 2011; 2013), or allow for some variability and examine this variability through moderator analysis. In accordance with other meta-analyses in educational research (Camilli et al., 2010; Fukkink & Lont, 2007;

Piasta & Wagner, 2010) and due to a limited amount of randomized studies, non-randomized intervention studies (quasi- and non-experimental) were included in the meta-analyses. Randomized and non-randomized (QED and non-experimental) studies could have been additionally analyzed separately to obtain more confidence on the magnitude of effects. However, a significant effect modifier could not be found for the randomization procedure for teachers, but moderator analysis indicated significant differences in contrast scores (EG vs. CG) and change scores (pre to posttest), at least for child outcomes.

A meta-analysis directly addresses the issue of generalization in its analysis and interpretation, and therefore particular characteristics concerning the subjects, situations, and procedures need to be coded as moderators. Further, to counteract the criticism of meta-analytic techniques – that they mix “apples and oranges” by aggregating the findings from studies with different designs, varying measures and informants – these variables were considered and tested as potential modifiers in this project. Several methodological variables have been discussed in psychological, behavioral and educational research as being meaningful effect modifiers (see Wilson & Lipsey, 2001). Meta-regressions indicated that the effect was moderated by research procedure such as pretest differences, reliability levels, score levels, contrasts and informants to a certain extent. Adjusting for methodological variables, professional development effects on quality and student learning remained substantial. Considering child care setting, effect estimates were relatively robust, except for the meta-analysis on language and literacy development where training was more effective in classrooms with disadvantaged children.

Further, *confidence intervals* describe the uncertainty inherent in aggregated effect size estimates. Effect size is known precisely when the confidence interval is very narrow (e.g., 0.60 to 0.70), but a wider interval (e.g., 0.63 to 0.90) can still be precise enough to make decisions about the usefulness of the intervention (Schünemann et al., 2008, p. 370). In the case of the meta-analyses conducted in this study, the 95 % confidence intervals show (albeit with some degree of imprecision) the utility of professional development to enhance quality ratings ($CI = 0.55$ to 0.82), language and literacy scores ($CI = 0.29$ to 0.47) and social, emotional and behavioral development ($CI = 0.31$ to 0.60). However, the estimates for the lower confidence intervals were all in the positive range and still indicated small to medium effects ($g > 0.20$).

The situation was different for the single studies. There were lower confidence intervals in the negative range for 6 treatments on quality improvements and 13 treatments on language and literacy development. The findings indicated a positive aggregated treatment effect (point estimate), but the lower confidence interval was negative. In this case, conclusions on the usefulness of these interventions and their effectiveness cannot be drawn with great confidence.

Another issue concerning the generalization of studies is the *unit of analysis* used to investigate the phenomena. As in primary studies, the unit of analysis (e.g., individual student, classroom, school) is the unit that allows for generalization. It can be noted that effect sizes tend to be larger if the unit of analysis is of a higher order, for example, school or district, compared with assessments at the student level (Hall et al., 1994). The units of analysis were coded, but generally all the quality ratings were conducted at classroom level¹⁹ and developmental outcomes assessed or rated per individual student²⁰.

The *generalization of findings from English and non-English language investigations* is controversial concerning study design features and divergent early childhood education and care systems. In the meta-analysis on in-service training impacts on language and literacy, 4 German investigations were included and in the meta-analysis on social, emotional and behavioral development there was one study from Thailand. Looking at sample characteristics, the question of comparability between children described as at risk from Germany and the USA arises, in particular concerning the degree of disadvantage, poverty and social deprivation. Disadvantaged children and families live in different circumstances due to divergent health care and social welfare systems. Another point is the use of different developmental domains and instruments, given that instruments have been extensively discussed as effect modifiers. Different concepts in the measurement of language and literacy exist, and at first glance it appears that the instruments used in North American studies are more fine-grained and more closely aligned to in-service interventions. The meta-analyses showed that instruments for some subdomains (e.g., concepts about print) are only used and available in English-speaking countries. The generalization of results across national boundaries is complicated by curricula and pedagogical traditions, pre-service education and formal requirements, age-homogenous vs. heterogeneous classrooms, structural requirements, quality of classrooms, and course of the day (Kuger, Sechtig, & Anders, 2012). Nevertheless, the *inclusion of non-English language trials* in meta-analyses has been discussed as a potential language bias. Excluding studies conducted and published in non-English speaking countries can increase bias and reduce the precision of findings (Walker, Hernandez, & Kattan, 2008). When 50 meta-analyses were re-analyzed with and without non-English language trials, most of the effect sizes changed to the limited degree of 5 % (Jüni et al., 2002), indicating that the results were not distorted due to the language mix. Interestingly, findings show that non-English language trials consist of fewer participants. This also appears in the meta-analysis on language and literacy development, in which studies conducted and published in Germany had smaller sample sizes than studies from English-speaking countries. The aggregated effects were small, but studies conducted and published in English-speaking coun-

¹⁹ Except for ECERS-R scores aggregated at center level from Japel (2009) and target child observation through the inCLASS from Morris et al. (2010)

²⁰ Except for the ratings of student response to teacher questions from Girolametto et al. (2007) and Lee (2010).

tries had a slightly larger impact ($g' = 0.39$; $SE = 0.048$) than German studies ($g' = 0.24$; $SE = 0.134$). In the other meta-analyses, all the studies were published in English. However, the study by Thanasetkorn (2009), with the most effective in-service program to foster social development, was carried out in Bangkok, Thailand.

3. *Effect size interpretation*

Effect sizes help to interpret the practical utility of an intervention. Effect size interpretation is challenging and is dependent on the context. In general, effect sizes illustrate the magnitude of a result, from which practical consequences can be deduced (Ellis, 2010). The interpretation of effects is important because research has 2 audiences – peers and non-specialists. Therefore, the *significance of an effect* has a different meaning for researchers and the broader public (Cooper, 2008). There are several recommendations that describe effects as significant, consequential or notable. Comparing the effects of medical treatments with psychological, educational and behavioral treatments, Lipsey and Wilson (1993) suggest that statistically modest values (0.10 or 0.20 SD's) found in psycho-educational treatments cannot be treated as obviously trivial (p. 1199). The WWC (2013) sets the bar for “substantively important” at $d = 0.25$ (p. 23). According to both definitions, in-service programs have notable effects on quality ratings ($g = 0.73$), language and literacy achievement ($g = 0.40$), and social, emotional and behavioral development ($g = 0.48$).

Another way to interpret the *magnitude of effects* is to refer to effect size conventions with cut-offs. Cut-offs classify effects into small, medium and large “like t-shirt sizes” (Ellis, 2010, p. 41). Despite the general criticism concerning the use of arbitrary benchmarks, there are variations in cut-off models, for example, the well-known model of Cohen (1988) considers an effect size of > 0.20 as small, > 0.50 as medium, and > 0.80 as large, while others interpret an effect of < 0.40 as small, > 0.40 as moderate, and > 0.70 as large (Ellis, 2010; Schünemann et al., 2008). Within this dissertation, the magnitudes of effects were discussed using the benchmarks of Cohen (1988). However, even Jacob Cohen himself argued that his conventions should be used “only when no better basis for estimating ES is available” (Cohen, 1988, p. 25). Decontextualized criteria for the interpretation of effect sizes are judged as unhelpful (Hedges, 2008). Critics call for alternative effect size interpretation that respects empirical benchmarks relevant to the intervention, population and outcome measures (Cooper, 2008; Hill et al., 2008). Hill and colleagues (2008) illustrate 3 types of empirical benchmarks: “(a) normative expectations for growth over time in student achievement, (b) policy-relevant gaps in student achievement by demographic group or school performance, and (c) effect size results from past research for similar interventions and target populations” (Hill et al., 2008, p. 172). In addition, Cooper (2008) suggests the translation of effect sizes into measures that are easier to understand for policy makers and practitioners, such as the

U₃-Index or binominal effect size display (BESD). The WWC (2013) translates effect sizes into an improvement index that helps readers to judge the practical importance of an intervention.

At the current state of research on child development in relation to educational interventions in center-based care, effects on student achievement are in the small to medium-sized range (Anderson et al. 2003; Blok et al., 2005; Camilli et al., 2010; Nelson, Westhues, & MacLeod, 2003; Preschool Curriculum Evaluation Research Initiative, 2008). The cost-benefit analysis by Heckman (2008) demonstrates that these small effects on child development in early childhood are sufficient for long-term benefits in adulthood and financial returns for society.

4. *Quality of research papers*

Another limitation concerns the *quality of research papers on primary studies*. The quality of papers and sufficient information are critical for 2 reasons: a) the inclusion of studies and b) the coding of effect modifiers and theoretical moderators. Overall, the authors of 44 rigorously carried out studies were contacted to receive statistical information that was missing in the article, for example, the exact sample size from EG and CG at pre- and posttest or the mean standard deviation to estimate Hedges' *g*. Further, information for several variables in professional development interventions (i.e., form, explicit components, amount, fidelity etc.), content, objectives and underlying didactic principles were missing. Often information was reported very ambiguously, so that the coders had difficulties in distinguishing between 2 categories. As an example of ambiguous reporting, it is difficult to discriminate between coaching and mentoring when reading the following sentence: "Support includes mentor teachers to coach teaching staff." (Barnett et al., 2008, p. 303) The precision and reliability of the coding suffer from such inconclusive information. For this reason, some rating categories on in-service interventions had poor inter-rater reliability values, and the coders had to discuss the findings, reading the paper once more to make final decision. Ambiguous information was also a problem concerning background information for trainees and providers, concepts and models for coaching, consulting or mentoring, and the quality and fidelity of in-service training implementation.

5.3. Implications

A growing body of evidence shows that teachers with academic degrees in education do not guarantee high-quality classrooms or the maximizing of children's academic gains (Early et al., 2006; 2007; Burchinal, Hyson, & Zaslow, 2008). Despite the fact that a certain level of teachers' qualifications (at least a college degree) contributes to a limited extent to classroom quality (Kelley & Camilli, 2007), teacher education alone is insufficient to transfer knowledge and skills to practice (Sheridan et al., 2009). In particular, teachers may benefit from background knowledge that pre-service education provides, but in-service training is important to ensure that the quality is achieved and sustained (Burchinal, Hyson, & Zaslow, 2008). Therefore, increasing the effectiveness of early childhood education requires a broad range of professional development activities (Early et al., 2007). In conclusion, quality in early childhood education and care and effects on student achievement can be improved when teachers receive carefully selected and implemented professional development (Burchinal, Hyson, & Zaslow, 2008).

5.3.1. Professional development research

A limited amount of rigorous research indicates that there is a *connection* between in-service professional development, early education practice and quality and student achievement. Meta-analyses demonstrated a link between in-service training and classroom quality and student gains in the areas of language, literacy and social, emotional, and behavioral development. Further, student learning is predicted by in-service training impacts on classroom quality based on a small amount of evidence. Considering the link from in-service training to student achievement as a three-step process, the evidence base is insufficient to make solid conclusions. There is a need for critical research to examine the process variables that promote changes in teachers' knowledge, attitudes, awareness and skills that result in teacher competencies. Further, research on process variables that promote the transfer from better skills into effective teaching and interventions that enhance student-teacher interactions and classroom quality is at an early stage of development. Last but not least, there is a lack of research that describes the path from professional development mediated by improved classroom practices to child learning. Several studies report professional development effects on child development, not knowing which changes in classroom interactions and practices made the change.

There is a need for critical research to *examine factors that contribute to in-service training effectiveness*. In line with other reviews on professional development research (e.g., Sheri-

dan et al., 2009; Snyder et al., 2012; Yoon et al., 2007), recommendations for professional development research concern the evaluation of characteristics of participants and providers and their roles and relationships, in-service training content, delivery, quality, and contextual aspects. In particular, the following aspects should be added to the research agenda to investigate potential modifiers:

- Identify key variables in in-service training delivery and their underlying didactic principles that modify outcomes at classroom and child level
- Discover the influence of the content of interventions
- Unpack the contribution of roles and relationships between trainer and trainee
- Uncover the (motivating) role of compensation and incentives
- Analyze the influence of collective participation and child care teams.

Potential effect modifiers can be analyzed in experimental studies that provide data. Further, secondary data analysis with correlational or regression techniques would be possible, if data on potential effect modifiers were investigated. Further, rigorous experimental investigations with variations in professional delivery and intensity (e.g., different amounts and duration, individual teachers vs. teams, traditional vs. modern formats) are needed to assess the effectiveness of training with greater precision. To unpack the contribution of incentives, motivation and relationships between trainees and trainers, mixed-method designs are required that combine experimental trials with qualitative methods. Differential analyses on in-service training effects for different groups of participants (e.g., teachers with high and low motivation or high and low values of self-efficacy) could elucidate the influence of the participants.

Another important aspect is *duration and time span*. Teachers' professional development is often criticized as being too short to provide adequate learning opportunities for them to integrate their new knowledge into practice (Penuel et al., 2007). Buysse, Winton, and Rous (2009) recommend that professional development should be sustained over time. Training transfer is defined as the maintenance of acquired knowledge and skills over time (Baldwin, Ford, & Blume, 2009). Conclusions on professional development impacts are almost exclusively based on findings from posttest measures illustrating short-term effects. A limited number of studies provide follow-up measures to depict and provide conclusions on time-lagged or fading effects. There are almost no studies with intermediate measures that offer information on teachers' learning processes.

To strengthen future research designs, random assignment would allow for stronger conclusions on causal inference. *Rigorous evaluation* requires the use of sophisticated data analyses to address the nesting of students within classrooms and schools. To detect in-service training effects more accurately, teacher, classroom, and student measures should be matched to the instructional topics presented in the professional development. Reliable measures with smaller measurement error, and which use standardized well-known scales,

will lead to stronger effect sizes. In the current meta-analyses, the equivalence of experimental and control teachers was one of the major concerns. Pretest differences were always powerful in predicting posttest values. Meta-analyses showed that small (non-significant) differences between groups in final outcome measures at pretest moderated the training effects. When differences between experimental and control groups are small, statistical adjustment of pretest differences works well with covariate, regression or HLM analyses. Further, propensity score matching is a valuable tool to reduce the problem of comparing similar students in dissimilar groups (Slavin, 2008).

For a better understanding of in-service training impacts on child development, further research should consider the influence of children's families and family practices. Additionally, alternative ways of interpreting effect sizes are needed that translate the magnitude of an effect into practical consequences, and that consider both the audiences of research – scientists and the broad public.

Last but not least, in line with the discussion by Snyder et al. (2012), several recommendations for improving reporting practices and the quality of study reports can be made. Researchers should clearly describe the form of professional development so that comparisons with other studies can be drawn with greater confidence. The review by Snyder et al. (2012) provides clear categories based on the framework from the NPDCI (2008). Further, it is necessary to specify process variables and active components, provide information about learners and contexts as well as providers, and report the amount, duration, rhythm and fidelity of in-service training interventions.

5.3.2. Professional development for early education teachers

The meta-analyses and other research summaries show that nearly the full spectrum of education and training interventions are available for early childhood teachers – from regional conferences, introductory workshops, online and blended learning formats, in-house and external courses or continuing education programs (Baumeister & Griesse, 2011; Beher & Walter, 2010; Snyder et al., 2012). Most of these interventions are informal, unsystematic and do not accumulate to obtain a credential (an agency or organization standard qualification) or credit units for formal education (Oberhuemer, 2012). Early childhood educators are most likely to be engaged in brief workshops or courses (Beher & Walter, 2012; Hamre & Hatfield, 2012). Such workshops are often ineffective in changing teacher competencies and practice (Garet et al., 2001). Overall, meta-analyses demonstrate that professional development can improve classroom quality and that this leads to beneficial effects on student achievement. Certainly, findings show, on the one hand, that some learning approaches are

more effective than others. On the other hand, results indicate that some in-service trainings are ineffective or can even cause negative consequences. In summary, five recommendations can be derived from this research synthesis on professional development effects.

1. Provide the appropriate training intensity to promote quality and student learning

In the USA and several European countries (e.g., U.K., Spain, and Portugal), *a certain amount of professional development per year* is required for early education teachers by state law (Eurydice, 2014; Oberhuemer, 2012; Whitebook et al., 2009a; 2009b). Additionally, a certain number of paid professional days per year is prescribed by state regulations, at least for K-12 teachers in the USA. In Germany, paid educational leave is established in some federal state regulations, but usually child care providers or centers set a fixed number of paid educational days per year (Beher & Walter, 2012; Peucker et al., 2010). In particular, approximately 5 in-service training days per year were offered to each early childhood practitioner in Germany (Beher & Walter, 2012). According to survey findings, early child care practitioners use predominately informal, short-term approaches, such as courses, workshops or conferences (Oberhuemer, 2012). For example, over 80 % of teachers in Germany participated in approaches with one to 3 training days and only 15 % in programs that ran for more than 3 days (Beher & Walter, 2012). Findings from the meta-analyses show that short-term workshop approaches on daily base are less effective. Approximately 30 hours of training are adequate to attain changes in single aspects of classroom practice that foster a specific developmental domain – either language and literacy or social behavior abilities. More training hours (> 45 hours) are suggested to improve classroom quality through in-service training meaningfully. Nevertheless, training intensity is always matched to content and training opportunities provided, but it is important to ensure that intensive training is available.

2. Use professional development components that ensure effective learning

Based on current evidence, *performance-based in-service programs* are more successful in improving classroom quality and student achievement. Performance-based in-service training provides specific skill training for on-the-job practice and incorporates field experience through performance-based assignments. Such in-service training approaches are content and domain-specific, and demonstrate target skills through modeling or videos and providing feedback to teachers. There are several programmatic recommendations in the literature which can serve as a starting point to establish quality standards for ongoing professional development at a local, state or national level. The recommendations point out the importance of ongoing professional development which is sustained over time, has specific objectives that are focused on practice and is coherent with learning standards and the classroom, and provides intensive learning opportunities with hands-on training for early ed-

education teachers (Buysse, Winton, & Rous, 2009; Deutsches Jugendinstitut, & Weiterbildungsinitiative Frühpädagogische Fachkräfte, 2014; NPDCI, 2008; U.S. Department of Education, 2002). At least for some subject areas (e.g., language promotion, inclusion, relationships with families), specific recommendations for content and performance-based approaches for in-service training providers exist in Germany (cf. Deutsches Jugendinstitut & Weiterbildungsinitiative Frühpädagogische Fachkräfte, 2011a; 2011b; 2013).

Most of the evaluated programs that were delivered to teachers were time and cost intensive and with extensive administration. Coaching was an effective approach to improve classroom quality, but it is probably one of the most cost-intensive professional development interventions. Hamre and Hatfield (2012) argue that in-service training is needed that is more *feasible, sustainable and affordable*. Prominent examples of programs to improve aspects of classroom quality are the 4-hour individual training program by Englund (2010) and 32-hour training course by Breffni (2011). Student achievement was also improved with low levels of teacher support. In particular, a 12-hour course (Cordell, 2010) or a half-day workshop with 4 follow-up sessions (Thanasetkorn, 2009) were enough to promote social skills. Buschmann et al. (2010b) and Girolametto et al. (2003) provided evidence that language and literacy skills can be fostered through short-term coursework and video feedback. This indicates that *video feedback* is a feasible tool to ensure teacher reflection and learning. A small but growing body of evidence shows that video feedback is very effective in changing teacher-student interactions and teaching behavior (cf. Fukkink, Trienekens, & Kramer, 2011). Evidence-based video feedback models for early childhood teachers to enhance teacher-student-interactions are available, for example, Video Interaction Guidance (Fukkink & Tavecchio, 2010), video supervision (Simon & Sachse, 2013) or micro-teaching (Beller, Merkens, & Preissing, 2007).

3. *Ensure quality of professional development*

In line with findings from other reviews (see Snyder et al., 2012), reporting on *the quality and fidelity of implemented professional development activities* was lacking. When fidelity was measured, providers completed checklists or attendance lists. Fidelity checklists require clear definitions of activities and a formal set of criteria for grading performance on a particular task. These formal criteria and activities in professional development can be summarized in a training manual or script. As a starting point to define professional development quality, in-service training activities and onsite models need to be specified. These specifications of in-service training components, principles and activities can help researchers and practitioners to understand the mechanism of professional development effects (see Sheridan et al., 2009).

Findings from the narrative review on the most effective in-service programs for promoting language and literacy abilities show that the *skills and training of in-service training providers* are critical in determining the effectiveness and quality of in-service programs. However, only a limited amount of studies reported information on the qualifications and training of providers, including formal education, experience in teaching in early childhood classrooms or coaching/consulting. When this information was provided, in the majority of cases, the providers were not experienced in providing onsite support (see Snyder et al., 2012). Several requirements for providers are proposed by the scientific community (e.g., Deutsches Jugendinstitut & Weiterbildungsinitiative Frühpädagogische Fachkräfte, 2014), but the current research base has not determined precisely what qualities providers should possess (Institute of Medicine & National Research Council of the National Academies, 2012; Whitebook et al., 2009b). The narrative review suggests that in-service training is more effective when it is manual-based and trainers are certified in the specific program. The training and certification of providers from highly effective programs consist of practical training, reliability of video ratings and behavior rehearsals (Buschmann et al., 2010b; Girolametto et al., 2003).

4. In-service training should cover relevant topics

The meta-analyses suggest that content-specific professional development approaches are needed that cover different developmental domains. The systematic literature search shows that there is a *lack of training approaches that foster the development of math and science* in the early years. In general, in-service training should be matched to standards for practice (Zaslow et al., 2010b). Early learning standards and state-regulated curricula set goals for the effectiveness of training in early childhood education. Further, they set demands for teachers to promote several education domains (language, math, science, art etc.) during the preschool years (Scott-Little et al., 2007). The promotion of several educational domains – programmatic or integrated – requires the provision of different in-service programs to enhance domain-specific teaching and interventions. The short-term offers currently provided are insufficient to accomplish this goal. This implies, on the one hand, that teachers should be required to participate in several in-service training programs to enhance quality and promote several developmental domains. On the other hand, it needs to be guaranteed that a variety of in-service programs will be available that cover relevant topics with sufficient amounts of training.

As mentioned earlier, in-service professional development has an essential role in increasing the effectiveness of early childhood education. Across countries, several efforts have been made to improve quality in early childhood settings. Most English-speaking countries implement Quality Improvement and Rating Systems at state or national levels (e.g., USA, England, Scotland etc.). In Germany, federal funds have been invested in a national quality initia-

tive (Nationale Qualitätsoffensive – NQI) to develop quality criteria, evaluation instruments, and quality management systems for early childhood education providers (BMFSFJ, 2003; 2004). According to a teacher survey, approximately 40 % of early childhood practitioners participate in quality circles regularly (Beher & Walter, 2012), but the survey also shows that there is a considerable demand (by more than 60 % of teachers) for training to improve quality. Improving process or interaction quality in classrooms is a broad and comprehensive training endeavor that requires more long-term and intensive professional development activities (cf. Hamre & Hatfield, 2012; Zaslow, 2009; Zaslow et al., 2010a; 2010b). Based on the meta-analysis that consisted only of studies from North America, it is suggested that in-service programs with 45 to 60 training hours can change quality sustainably. However, there is a lack of studies investigating quality improvement training in Europe. Findings from the pilot study by Erning (2003) demonstrate that quality in early childhood classrooms in Germany can be improved through a multi-year in-service program that consists of a quality rating and feedback on the ECERS and coursework to elaborate knowledge on child care quality. This model is analogous to the approaches used in the QRIS (see Boller et al., 2010). The experience from quality improvement initiatives in the United States which offer intensive training opportunities for teachers looks promising. These training principles and evaluation concepts can easily be translated into other countries, because the well-known quality rating scales have already been made available, translated, and adapted.

5. *Systematize professional development efforts*

Considering the unsystematic provision of ongoing professional development, *system level components* that ensure appropriate selection and ongoing training should become the focus of attention. However, several policy factors should be considered that impede or accelerate the process towards effective professional development. At the state level, there are divergent requirements for early childhood staff and regulations for training qualifications (Autorengruppe Bildungsberichtserstattung, 2014). Many individuals, organizations and agencies offer ongoing professional development opportunities, but professional development standards, for example, the number of training days, are set autonomously by administrators and local welfare agencies or centers (Peucker et al., 2010). The existence of several professional development operators may lead to divergent quality in in-service training programs. Governments and the scientific community have developed *standards for the quality of ongoing professional development* for teachers (Deutsches Jugendinstitut & Weiterbildungsinitiative Frühpädagogische Fachkräfte, 2014; U.S. Department of Education, 2006), but so far most of them are non-compulsory. As regards their delivery, high-quality professional development programs should be “sustained, intensive, and classroom-focused and are not short-term workshops” (U.S. Department of Education, 2006, p. 1). In-service programs improve

teachers' knowledge and skills to promote children's learning, advance teachers' understanding of effective evidence-based instructional strategies, and enhance classroom management skills (U.S. Department of Education, 2006). Some first efforts have been made to define quality standards and indicators for the evaluation of professional development quality (Expertengruppe Berufsbegleitende Weiterbildung, 2013). The findings from the narrative review suggest that more transparency on content, principles of learning, and the training of providers is needed. Transparency and quality can be achieved through a licensing and quality management procedure for providers and their activities (von Hippel & Grimm, 2010). In the case of the K-12 system in the United States, there are ongoing professional requirements at state level. Lists of certified providers and activities are released on the state education department website to ensure the quality of in-service programs. Public K-12 school teachers in the United States are required to have a record of individual ongoing professional development. Such an ongoing professional development system has already been established in Germany for medical practitioners, administrated by the medical council (Ärztekammer), but not for school or preschool teachers. Such an ongoing professional development system that regulated and certified professional development provision and provided financial funds would help to systematize in-service training interventions and ensure high-quality training opportunities for early childhood practitioners.

Most of the provided in-service training is not cumulative and is diverse in quality. Strategies for continuing professional development are needed that ensure the participation of teachers and that are consecutive and result in credentials or educational credits. There is a growing need for a *credit transfer (and accumulation) system* for ongoing professional development (cf. Expertengruppe Berufsbegleitende Weiterbildung, 2013). The Council of the European Union (2012) has recommended the development of a system that enables the accumulation of competencies based on non-formal and informal education. Across nations, policies have been developed for standards on early childhood educator competencies for pre- and in-service education (cf. California Department of Education & First 5, 2011; Gebrande, 2011). In Germany, there are individual and generic credit transfer systems that allow early childhood practitioners to enroll in higher education programs, but more feasible systems are needed that reduce the amount of personal capacity for the individual accumulation of credits (Fröhlich-Gildhoff & Röser, 2013; Kruse, 2012).

Findings from the narrative review point out that in-service programs were very effective in improving early child care quality, when *financial compensations* were offered. Financial compensation consists of personal stipends for teachers that cover tuition fees, and funds for expenses for release time for the child care center. The results from a teacher survey indicate that the high costs of in-service training and staff shortages prevent teachers from participating in intensive training (Beher & Walter, 2012). To ensure participation in intensive

professional development approaches, financial compensation should be seriously considered. In the US, a national comprehensive scholarship program called T.E.A.C.H. has been established to promote teachers to take coursework to obtain a credential or a higher education degree (cf. Miller & Bogatova, 2009). Such a model would be helpful to encourage teachers to take classes.

An improved learning environment and teaching practice in a systematic and state-wide approach is the aim of *induction programs* for novice teachers. Despite their background knowledge from formal education, researchers consider their ongoing training and feedback, especially in interactions with students, as the key to ensure high quality in early childhood education classrooms (Burchinal, Hyson, & Zaslow, 2008). In K-12 education, it is assumed that new teachers need a period of support in order to develop into an effective practitioner, but this concept has not been established in early child care (Whitebook et al., 2009a, p. 5). An inductive program is most effective when it includes guidance and assistance (e.g., models teacher practices and provides opportunities for feedback and reflection) by a well-trained and skilled coach or mentor (Whitebook et al., 2009a; 2009b).

In conclusion, the research on professional development shows that there are several programs that have been proven to work. Nevertheless, most in-service training programs are short-term, and most funds are spent on workshops, despite the fact that brief workshops do not promote meaningful changes in teaching practice and child development (Behr & Walter, 2010; 2012; Hamre & Hatfield, 2012). The allocation of investment in in-service training should be aligned with evidence-based practices. Evidence-based training formats should be brought to scale instead of non-evident workshops. Scaling up and the allocation of investments require changes at policy level and funding for research on the implementation of effective programs at scale.

To sum up, the current meta-analyses provide scientific evidence that the early childhood field has a unique chance to invest in the development of early education practitioners through in-service training. This investment has a direct impact on early education quality and student achievement.

6. Literature

* Studies included in the meta-analyses are marked with an asterisk

- Abbott, M., Atwater, J., Lee, Y., & Edwards, L. (2011). A data-driven preschool professional development model for literacy and oral language instruction. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 229-245.
- Abrams, & Company. (2000). *Let's begin with the letter people*. Waterbury, CT: Abrams & Company.
- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- *Ahrens, R. (2009). *Training childcare providers to improve interactions with children: A comparison of in-service training versus direct modeling of desired behaviors*. Unpublished master thesis. University of Nebraska at Omaha. Omaha.
- Akobeng, A. K. (2005a). Understanding systematic reviews and meta-analysis. *Archives in Disease in Childhood*, 90(8), 845-848.
- Akobeng, A. K. (2005b). Principles of evidence based medicine. *Archives in Disease in Childhood*, 90(8), 837-840.
- *Algozzine, B., Babb, J., Algozzine, K., Mraz, M., Kissel, B., Spano, S., et al. (2011). Classroom effects of an early childhood educator professional development partnership. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 246-262.
- Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, 3(4), 385-416.
- American Psychology Association. (2009). *Publication manual of the American Psychology Association. 6th Edition*. Washington, D.C.: American Psychology Association.
- American Psychology Association's Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, 61(4), 839-851.
- Anders, Y. (2012). Modelle professioneller Kompetenz für frühpädagogische Fachkräfte. Aktueller Stand und ihr Bezug zur Professionalisierung. Expertise zum Gutachten "Professionalisierung in der Frühpädagogik". Retrieved March 22nd, 2013, from http://www.aktionsrat-bildung.de/fileadmin/Dokumente/Expertise_Modelle_professioneller_Kompetenzen.pdf
- Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., & Normand, J. (2003). The effectiveness of early childhood development programs. A systematic review. *American Journal of Preventive Medicine*, 24(3), 32-46.
- Arkoosh, M. K., Derby, K. M., Wacker, D. P., Berg, W., McLaughlin, T. F., & Baretto, A. (2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification*, 31(6), 880-895.
- Armstrong, K., Cusumano, D. L., Todd, M., & Cohen, R. (2008). Literacy training for early childhood providers: Changes in knowledge, beliefs, and instructional practices. *Journal of Early Childhood Teacher Education*, 29(5), 297-308.
- Arnold, D. H., Fischer, P. H., Doctoroff, G. L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology*, 94(4), 762-770.
- *Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007). An evaluation of curriculum, setting, and mentoring on the performance of children enrolled in pre-kindergarten. *Reading and Writing*, 20(5), 463-494.
- Autorengruppe Bildungsberichtserstattung. (2012). *Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebenslauf*. Bielefeld: Bertelsmann.
- Autorengruppe Bildungsberichtserstattung. (2014). *Bildung in Deutschland 2014. Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen*. Bielefeld: Bertelsmann.

- *Baker, S., & Smith, S. (1999). Starting off on the right foot: The influence of four principles of professional development in improving literacy instruction in two kindergarten programs. *Learning Disabilities Research and Practice*, 14(4), 239-253.
- *Baker-Henningham, H., Scott, S., Jones, K., & Walker, S. (2012). Reducing child conduct problems and promoting social skills in a middle-income country: Cluster randomised controlled trial. *The British Journal of Psychiatry*, 201(2), 101-108.
- Baker-Henningham, H., Walker, S., Powell, C., & Meeks Gardner, J. M. (2009). A pilot study of the Incredible Years teacher training programme and a curriculum unit on social and emotional skills in community pre-schools in Jamaica. *Child: Care, Health and Development*, 35(5), 624-631.
- Baldwin, T. T., Ford, J. K., & Blume, B. D. (2009). Transfer of training 1988-2008: An updated review and new agenda for future research. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (pp. 41-70). Chichester, UK: Wiley.
- Baldwin, T. T., & Ford, K. J. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41(1), 63-105.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612-637.
- Barnett, S. W. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *Future of Children*, 5(3), 25-50.
- Barnett, S. W. (2004). Better teachers, better preschools: Student achievement linked to teacher qualifications. *NIEER Policy Brief*, 2, 1-12.
- Barnett, S. W., & Frede, E. (2010). The promise of preschool: Why we need early education for all. *American Educator*, 34(1), 21-29.
- *Barnett, S. W., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., et al. (2008). Educational effects of the tools of the mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, 23(3), 299-313.
- Baumeister, K., & Grieser, A. (2011). Berufsbegleitende Fort- und Weiterbildung frühpädagogischer Fachkräfte – Analyse der Programmangebote. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Studien, Band 10. Retrieved May, 5th, 2013, from http://www.weiterbildungsinitiative.de/uploads/media/Baumeister_Grieser_pdf.pdf
- Baumeister, R. F. & Leary, M. R. (1997). Writing narrative literature reviews. *Review of General Psychology*, 1(3), 311-320.
- Beelmann, A. (2004). Abflachung ökologischer Übergänge: Vom Kindergarten in die Schule. in G. W. Lauth, M. Grünke & J. C. Brunstein (Eds.), *Interventionen bei Lernstörungen. Förderung, Training und Therapie in der Praxis* (pp. 432-445). Göttingen: Hogrefe.
- Beelmann, A. (2014). Möglichkeiten und Grenzen systematischer Evidenzkumulation durch Forschungssynthese in der Bildungsforschung. *Zeitschrift für Erziehungswissenschaften*, 17(Supplement), 55-78.
- Behr, K., & Walter, M. (2010). Zehn Fragen - Zehn Antworten zur Fort- und Weiterbildungslandschaft für frühpädagogische Fachkräfte. Werkstattbericht aus einer bundesweiten Befragung von Weiterbildungsanbietern. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Studien, Band 6. Retrieved April, 16th, 2013, from http://www.weiterbildungsinitiative.de/uploads/media/WiFF_Studie_6_Walter_Behr_Internet.pdf
- Behr, K., & Walter, M. (2012). Qualifikationen und Weiterbildung frühpädagogischer Fachkräfte. Bundesweite Befragung von Einrichtungsleitungen und Fachkräften in Kindertageseinrichtungen: Zehn Fragen - Zehn Antworten. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Studien, Band 15. Retrieved June 26th, 2013, from http://www.weiterbildungsinitiative.de/uploads/media/Studie_BehrWalter.pdf
- Bell, B. S., & Kozlowski, S. W. J. (2002). Adaptive guidance: Enhancing self-regulation, knowledge, and performance in technology-based training. *Personnel Psychology*, 55(2), 267-306.

- *Beller, K., Merckens, H., & Preissing, C. (2007). *Erzieherqualifizierung zur Erhöhung des sprachlichen Anregungsniveaus in Tageseinrichtungen für Kinder - eine Interventionsstudie*. Berlin: Internationale Akademie für innovative Pädagogik Psychologie und Ökonomie gGmbH (INA) an der Freien Universität Berlin.
- *Beller, S., & Beller, E. K. (2009). *Abschlussbericht des Projekts. Systematische sprachliche Anregung im Kindergartenalltag zur Erhöhung der Bildungschancen 4- und 5-jähriger Kinder aus sozial schwachen und Migrantenfamilien - ein Modell der pädagogischen Intervention*. Berlin: Freie Universität Berlin.
- Berg, D., & Tisdale, T. (2004). Verhaltensauffälligkeiten bei Grundschulkindern. Eine elektronische Publikation einer epidemiologischen Studie. Retrieved August 28th, 2014, from <http://www.ub.uni-bamberg.de/elib/volltexte/2004/8/bergtisdale1.pdf>
- Berufsbildungswerk Leipzig. (2011). Abschlussbericht. Landesmodellprojekt. „Sprache fördern“. Erprobung und Multiplikation von Methoden der Sprachförderung in Kindertagesstätten. Laufzeit: 2007-2011. Retrieved June 26th, 2013, from http://www bbw-leipzig.de/fileadmin/user_upload/1_Gruppe/Downloads/Abschlussbericht_Sprache_foerdern.pdf
- Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership*, 57(8), 28-33.
- Bishop, D. V. M. (2001). Genetic influences on language impairment and literacy problems in children: Same or different? *Journal of Child Psychology and Psychiatry*, 42(2), 189-198.
- Bishop, D. V. M., Adams, C. V., & Frazier Norbury, C. (2004). Using nonword repetition to distinguish genetic and environmental influences on early literacy development. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 129B(1), 94-96.
- Blank, R. K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement. How meta analysis provides scientific evidence useful to education leaders*. Washington, D.C.: Council of Chief State School Officers (CCSSO).
- Blok, H., Fukkink, R. G., Gebhardt, E. C., & Leseman, P. M. (2005). The relevance of delivery mode and other programme characteristics for the effectiveness of early childhood intervention. *International Journal of Behavioral Development*, 29(1), 35-47.
- *Bloom, P. J. & Sheerer, M. (1992). The effects of leadership training on child care program quality. *Early Childhood Research Quarterly*, 7(4), 579-594.
- Blume, B. D., Ford, K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, 36(4), 1065-1105.
- Bodrova, E., & Leong, D. J. (2007). *Tools of the Mind: The Vygotskian approach to early childhood education* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
- *Boller, K., Del Grosso, P., Blair, R., Jolly, Y., Fortson, K., Paulsell, D., et al. (2010). *The Seeds to Success modified field test: Findings from the impact and implementation studies*. Princeton, NJ: Mathematica Policy Research, Inc.
- Bonate, P. L. (2000). *Analysis of pretest posttest design*. Boca Raton, FL: Chapman & Hall.
- Borenstein, M. (2009). Effect size for continuous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis. Second edition* (pp. 221-235). New York, NY: Russell Sage Foundation.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-analysis (CMA) Version 2*. Englewood, NJ: Biostat.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaften*. Heidelberg: Springer.
- *Bos, C. S., Mather, N., Friedman Narr, R., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. *Learning Disabilities Research & Practice* 14(4), 227-238.

- Brady, S. (1997). Ability to encode phonological representations: An underlying difficulty of poor readers. In B. A. Blachman (Ed.), *Foundation of reading acquisition and dyslexia: Implications for early intervention* (pp. 21-47). Hillsdale, NJ: Lawrence Erlbaum.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C.: National Academy Press.
- Bredekamp, S. (1986). The reliability and validity of the Early Childhood Classroom Observation Scale for accrediting early childhood programs. *Early Childhood Research Quarterly*, 1(1), 103-118.
- *Breffni, L. (2011). Impact of curriculum training on state-funded prekindergarten teachers' knowledge, beliefs, and practices. *Journal of Early Childhood Teacher Education*, 32(2), 176-193.
- *Brinks, R. (2007). *Intensive professional development in early literacy instruction for pre-school teachers*. Unpublished dissertation, Western Michigan University, Kalamazoo.
- *Brown, B. L. A. (2008). *The effects of consultation and professional development on pre-literacy skills of preschool children from low-income backgrounds*. Unpublished master thesis, California State University, Long Beach.
- Brown, J. R., Knoche, L. L., Edwards, C. P., & Sheridan, S. M. (2009). Professional development to support parent engagement: A case study of early childhood practitioners. *Early Education and Development*, 20(3), 482-506.
- Brown, T. (2005). Effectiveness of distal and proximal goals as transfer of training intervention: A field experiment. *Human Resource Development Quarterly*, 16(3), 369-387.
- *Bryant, D., Wesley, P., Burchinal, M., Sidersis, J., Taylor, K., Fenson, C., et al. (2009). *The QUINCE-PFI Study. An evaluation of a promising model for child care provider training*. Chapel Hill, NC: University of North Carolina at Chapel Hill, FPG Child Development Institute.
- Bryk, A. S., & Raudenbush. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.
- Bundesministerium für Familie, Senioren, Frauen und Jugend [BMFSFJ]. (2003). *Die Förderung nach Bildungsqualität. Auf den Anfang kommt es an! Perspektiven zur Weiterentwicklung des Systems der Tageseinrichtungen für Kinder in Deutschland*. Weinheim: Cornelsen Verlag.
- Bundesministerium für Familie, Senioren, Frauen und Jugend. (2004). Nationale Qualitätsinitiative im System der Tageseinrichtung für Kinder. Angebote zur Umsetzung der Ergebnisse. Retrieved December 14th, 2013, from <http://www.kita-bildungsserver.de/downloads/download-starten/?did=420>
- Bundesministerium für Familien, Senioren, Frauen und Jugend. (2012). Programmüberblick - Offensive Frühe Chancen: Schwerpunkt-Kitas Sprache & Integration. Retrieved December 27th, 2012, from http://www.fruehe-chancen.de/files/schwerpunkt_kitas/application/pdf/programmueberblick.pdf
- Burchinal, M. (2010). *Differentiating among measures of quality: Key characteristics and their coverage in existing measures*. OPRE Research-to-Policy, Research-to-Practice Brief OPRE 2011-10b. Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Burchinal, M., Hyson, M., & Zaslow, M. (2008). Competencies and credentials for early childhood educators: What do we know and what do we need to know? NHSA Dialog, 11(1). Retrieved February 25th, 2013, from http://www.nhsa.org/files/static_page_files/9FFE3481-1D09-3519-ADDD70A596F4B2FF/Dialog_Briefs_v11no1.pdf
- Burchinal, M. P. (2008). How measurement error affect the interpretation and understanding of effect size. *Child Development Perspectives*, 2(3), 178-180.
- *Burchinal, M. R., Cryer, D., Clifford, R. M., & Howes, C. (2002). Caregiver training and classroom quality in child care centers. *Applied Developmental Science*, 6(1), 2-11.

- Burchinal, M. R., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies on early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 11-31). Baltimore, MD: Brookes.
- Burchinal, M. R., Roberts, J. E., Nabors, L. A., & Bryant, D. M. (1996). Quality of center child care and infant cognitive and language development. *Child Development*, 67(2), 606–620.
- Burchinal, M. R., Roberts, J. E., Riggins, R., Zeisel, S. A., Neebe, E., & Bryant, D. (2000). Relating quality of center child care to early cognitive and language development longitudinally. *Child Development*, 71(2), 606–620.
- Burke, L. A. (1997). Improving positive transfer: A test of relapse prevention training on transfer outcomes. *Human Resource Development Quarterly*, 8(2), 115-128.
- Burke, L. A., & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *American Journal of Public Health*, 6(3), 263-296.
- Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Salvador, R. O., & Islam, G. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health*, 96(2), 315-324.
- Buschmann, A., Jooss, B., Simon, S., & Sachse, S. (2010a). Alltagsintegrierte Sprachförderung in Krippe und Kindergarten. Das "Heidelberger Trainingsprogramm". Ein sprachbasiertes Interaktionstraining für den Frühbereich. *Interdisziplinär*, 18(1), 84-95.
- *Buschmann, A., Simon, S., Jooss, B., & Sachse, S. (2010b). Ein sprachbasiertes Interaktionstraining für Erzieherinnen ("Heidelberger Trainingsprogramm") zur alltagsintegrierten Sprachförderung in Krippe und Kindergarten - Konzept und Evaluation. In K. Fröhlich-Gildhoff, I. Nentwig-Gesemann & P. Strehmel (Eds.), *Forschung in der Frühpädagogik III Schwerpunkt: Sprachentwicklung & Sprachförderung. Materialien zur Frühpädagogik. Band 5* (pp. 107-133). Freiburg: FEL Verlag.
- *Buysse, V., Castro, D. C., & Peisner-Feinberg, E. (2010). Effects of a professional development program on classroom practices and outcomes for latino dual language learners. *Early Childhood Research Quarterly*, 25(2), 194-206.
- Buysse, V., Winton, P. J., & Rous, B. (2009). Reaching consensus on a definition of professional development for the early childhood field. *Topics in Early Childhood Special Education* 28(4), 235-243.
- California Department of Education, & First 5. (2011). California early childhood educator competencies. Retrieved June 14th, 2013, from <http://www.cde.ca.gov/sp/cd/re/documents/ececompetencies2011.pdf>
- Camilli, G., Vargas, S., Ryan, S., & Barnett, S. W. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, 112(2), 579-620.
- Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching children to read: The fragile link between science and federal education policy. *Education Policy Analysis Archives*, 11(15), 1-51.
- *Campbell, P. H., & Milbourne, S. A. (2005). Improving the quality of infant? Toddler care through professional development. *Topics in Early Childhood Special Education*, 25(1), 3-14.
- Campbell, S. B., Spieker, S., Burchinal, M., & Poe, M. D. (2006). Trajectories of aggression from toddlerhood to age 9 predict academic and social functioning through age 12. *Journal of Child Psychology & Psychiatry*, 47(8), 791-800.
- Campbell Collaboration. (n.d.). Campbell Collaboration systematic reviews: Policies and guidelines. Retrieved June 15th, 2012, from http://www.campbellcollaboration.org/artman2/uploads/1/C2_Protocols_guidelines_v1.pdf

- *Cassidy, D. J., Buell, M. J., Pugh-Hoesse, S., & Russell, S. (1995). The effect of education on child care teachers' beliefs and classroom quality: Year one evaluation of the TEACH early childhood associate degree scholarship program. *Early Childhood Research Quarterly*, 10(2), 171-183.
- Chard, D. J., & Dickson, S. V. (1999). Phonological awareness: Instructional and Assessment Guidelines. *Intervention in School and Clinic*, 34(5), 261-270.
- Cheng, E. W. L., & Ho, D. C. K. (2001). A review of transfer of training studies in the past decade. *Personnel Review*, 30(1), 102-118.
- Chiaburu, D. S., & Marinova, S. V. (2005). What predicts skill transfer? An exploratory study of goal orientation, training self-efficacy and organizational supports. *International Journal of Training and Development*, 9(2), 110-123.
- *Clancy-Menchetti, J. L. (2006). *Early literacy professional development: Exploring the effects of mentoring for preschool teachers*. Unpublished dissertation, The Florida State University, Tallahassee.
- Clarke-Stewart, K. A., Vandell, D. L., Burchinal, M., O'Brien, M., & McCartney, K. (2002). Do regulable features of child-care homes affect children's development? *Early Childhood Research Quarterly*, 17(1), 52-86.
- Cohen, J. (1988). *Statistical power analysis for the behavioral science (2nd Edition)*. Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, L. E., Kramer-Vida, L., & Frye, N. (2012). Using dialogic reading as professional development to improve students' english and spanish vocabulary. *NHSA Dialog: A Research-to-Practice Journal of the Early Childhood Field*, 15(1), 59-80.
- Coie, J. D., & Kupersmidt, J. B. (1983). A behavioral analysis of emerging social status in boys' groups. *Child Development*, 54(6), 1400-1416.
- Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011). Correlation coefficients by student achievement level. Retrieved December 11th, 2012, from <http://ies.ed.gov/ncee/pubs/20114033/pdf/20114033.pdf>
- Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5), 678-707.
- Connors-Tadros, L., & Carlson, B. C. (2011). Integrating quality rating systems and professional development systems in early childhood. In C. Howes & R. C. Pianta (Eds.), *Foundation for teaching excellence. Connecting early childhood quality rating, professional development and competency systems in states* (pp. 25-46). Baltimore, MD: Brookes.
- Cooper, H. (2008). The search for meaningful ways to express the effects of intervention. *Child Development Perspectives*, 2(3), 181-186.
- *Cordell, M. (2010). *The impact of technical coaching in emergent literacy and language curriculum on a Head Start teacher, staff members, english language learners, and hispanic family literacy practices*. Unpublished dissertation, Graduate School of Texas A&M University, Commerce.
- Council of the European Union (2012). Council recommendation of 20 December 2012 on the validation of non-formal and informal learning. Retrieved October 25th, 2013 from [http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32012H1222\(01\)&from=DE](http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32012H1222(01)&from=DE)
- Crick, N. R., Casas, J. F., & Mosher, M. (1997). Relational and overt aggression in preschool. *Developmental Psychology*, 33(4), 579-599.
- *Cunningham, L. C. (2007). *Toward better early childhood education: Links between professional development and language and literacy practices*. Unpublished dissertation, The University of Michigan, Ann Arbor.
- Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., et al. (2009). The relations of observed pre-K classroom quality profiles to children's achievement and social competence. *Early Education & Development*, 20(2), 346-372.

- *Cusumano, D. L., Armstrong, K., Cohen, R., & Todd, M. (2006). Indirect impact: How early childhood educator training and coaching impacted the acquisition of literacy skills in preschool students. *Journal of Early Childhood Teacher Education*, 27(4), 363-377.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23-45.
- Decker, P. J. (1980). Effects of symbolic coding and rehearsal in behavior-modeling training. *Journal of Applied Psychology*, 65(6), 627-634.
- Decker, P. J. (1982). The enhancement of behavior modeling training of supervisory skills by the inclusion of retention processes. *Personnel Psychology*, 35(2), 323-332.
- Decker, P. J., & Nathan, B. R. (1985). *Behavior modeling training: Principles and applications*. New York, NY: Praeger.
- Deeks, J. J. H., Higgins, J. P. T., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 243-296). Chichester, UK: Wiley-Blackwell.
- Demchak, M., Kontos, S., & Neisworth, J. T. (1992). Using a pyramid model to teach behavior management procedures to childcare providers. *Topics in Early Childhood Special Education* 12(4), 458-477.
- Department for Education. (2012). Statutory framework for the Early Years Foundation Stage. Setting the standards for learning, development and care for children birth to five. Retrieved October 18th, 2013, from http://planipolis.iiep.unesco.org/upload/UK/UK_EYFS_Statutory_Framework.pdf
- Deutsches Jugendinstitut, & Weiterbildungsinitiative Frühpädagogische Fachkräfte. (2011a). Sprachliche Bildung. Grundlagen für die kompetenzorientierte Weiterbildung. WiFF Wegweiser Weiterbildung, Band 1. München. Retrieved December 16th, 2012, from http://www.weiterbildungsinitiative.de/uploads/media/WiFF_WegweiserWeiterbildung_1_SprachlicheBildung_Internet.pdf
- Deutsches Jugendinstitut, & Weiterbildungsinitiative Frühpädagogische Fachkräfte. (2011b). Zusammenarbeit mit Eltern. Grundlagen für die kompetenzorientierte Weiterbildung. WiFF Wegweiser Weiterbildung, Band 3. München. Retrieved December 16th, 2012, from http://www.weiterbildungsinitiative.de/uploads/media/WiFF_Wegweiser_3_Zusammenarbeit_mit_Eltern_Internet.pdf
- Deutsches Jugendinstitut, & Weiterbildungsinitiative Frühpädagogische Fachkräfte. (2013). Inklusion – Kinder mit Behinderung. Grundlagen für die kompetenzorientierte Weiterbildung. WiFF Wegweiser Weiterbildung, Band 6. München. Retrieved March 15th, 2014, from http://www.weiterbildungsinitiative.de/uploads/media/WW6_Inklusion_Kinder_mit_Behinderung.pdf
- Deutsches Jugendinstitut, & Weiterbildungsinitiative Frühpädagogische Fachkräfte. (2014). Kompetenzorientierte Gestaltung von Weiterbildung. Grundlagen für die Frühpädagogik. WiFF Wegweiser Weiterbildung, Band 7. Retrieved June 25th, 2014, from http://www.weiterbildungsinitiative.de/uploads/media/WiFF_Wegweiser_7_Kompetenzorientierte_Weiterbildung.pdf
- Dickinson, D. K., & Neuman, S. B. (2006). *Handbook of early literacy research*. Volume 2. New York, NY: Guilford.
- Dickinson, D. K., & Brady, J. P. (2006). Toward effective support for language and literacy through professional development. In M. Zaslow & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (pp. 141-170). Baltimore, MD: Brookes.
- *Dickinson, D. K., & Caswell, L. (2007). Building support for language and early literacy in preschool classrooms through in-service professional development: Effects of the literacy environment enrichment program (LEEP). *Early Childhood Research Quarterly*, 22(2), 243-260.

- Doane, S. M., Alderton, D. L., Sohn, Y. W., & Pellegrino, J. W. (1996). Acquisition and transfer of skilled performance: Are visual discrimination skills stimulus specific? *Journal of Experimental Psychology: Human Perception and Performance*, 22(5), 1218-1248.
- Doane, S. M., Sohn, Y. W., & Schreiber, B. (1999). The role of processing strategies in the acquisition and transfer of a cognitive skill. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5), 1390-1410.
- Dobbs-Oates, J., Kaderavek, J. N., Guo, Y., & Justice, L. M. (2011). Effective behavior management in preschool classrooms and children's task orientation: Enhancing emergent literacy and language development. *Early Childhood Research Quarterly*, 26(4), 420-429.
- Dodge, D. T., Colker, L., & Heroman, C. (2008). *The Creative Curriculum for preschool (College Edition)*. Washington, D.C.: Teaching Strategies Inc.
- Dodge, K. A. (1983). Behavioral antecedents of peer social status. *Child Development*, 54(6), 1386-1399.
- Dominguez, X., Vitiello, V. E., Fuccillo, J. M., Greenfield, D. B., & Bulotsky-Shearer, R. J. (2011). The role of context in preschool learning: A multilevel examination of the contribution of context-specific problem behaviors classroom process quality to low-income children's approaches to learning. *Journal of School Psychology*, 49(2), 175-195.
- *Domitrovich, C. E., Cortes, R. C., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool "PATHS" curriculum. *The Journal of Primary Prevention*, 28(2), 67-91.
- *Domitrovich, C. E., Gest, S. D., Gill, S., Bierman, K. L., Welsh, J. A., & Jones, D. (2009). Fostering high-quality teaching with an enriched curriculum and professional development support: The Head Start REDI program. *American Educational Research Journal*, 46(2), 567-597.
- Domitrovich, C. E., Gest, S. D., Jones, D., Sukhdepp, G., & Sanford DeRousie, R. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, 25(3), 284-298.
- Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interaction. *Early Childhood Research Quarterly*, 25(1), 1-16.
- Downer, J. T., López, M. L., Grimm, K., Hamagami, A., Pianta, R. C., & Howes, C. (2011a). Observations of teacher-child interactions in classrooms serving latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, 27(1), 1390-1410.
- *Downer, J., Pianta, R., Fan, X., Hamre, B., Mashburn, A., Justice, L. (2011b). Effects of web-mediated teacher professional development on the language and literacy skills of children enrolled in prekindergarten programs. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 189-212.
- Driscoll, K. C., Wang, L., Mashburn, A. J., & Pianta, R. C. (2011). Fostering supportive teacher-child relationships: Intervention implementation in a state-funded preschool program. *Early Education and Development*, 22(4), 593-619.
- Dubs, R. (1990). Lernprozesse in Unternehmungen beschleunigen – Zur Transferproblematik in Unternehmungen. *Die Unternehmung*, 44(3), 154-163.
- Dunst, C. J., & Raab, M. (2010). Practitioners' self-evaluations of contrasting types of professional development. *Journal of Early Intervention*, 32(4), 239-254.
- Dunst, C. J., & Trivette, C. M. (2011). Evidence-based strategies for training adults to use assistive technology and adaptations [Electronic Version]. *Research Brief*, 5(1), 1-8. Retrieved June 6th, 2013 from http://tnt.asu.edu/files/TotsNTech_ResearchBrief_2011.pdf.
- Dunst, C. J., Trivette, C. M., & Hamby, D. W. (2010). Meta-analysis of the effectiveness of four adult learning methods and strategies: Supplement tables and references. *International Journal of Continuing Education and Lifelong Learning*, 3(1), 91-112.

- Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., et al. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly*, 21(2), 174-195.
- Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D.-M. C., Crawford, G. M., et al. (2010). How do pre-kindergartners spend their time? Gender, ethnicity and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, 25(2), 177-193.
- Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*, 78(2), 558-580.
- Ebert, S., & Weinert, S. (2013). Predicting reading literacy in primary school: The contribution of various language indicators in preschool. In M. Pfof, C. Artelt & S. Weinert (Eds.), *The development of reading literacy from early childhood to adolescence: Empirical findings from the Bamberg BiKS longitudinal studies* (pp. 93-149). Bamberg: University of Bamberg Press.
- *Edgar, T. (2008). *The impact of a research-based professional development project on the language and literacy environment of pre-kindergarten teachers' classrooms*. Unpublished dissertation, University of Houston, Houston.
- Egger, M., Smith, G. D., Phillips, A. N. (1997a). Meta-analysis. Principles and procedures. *British Medical Journal*, 315(7121), 1533-1537.
- Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997b). Language bias in randomized controlled trials published in English and German. *The Lancet*, 350(9074), 326-329.
- Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997c). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109), 629-634.
- Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.
- Elmore, R. F. (2002). *Bridging the gap between standards and achievement. The imperative for professional development in education*. Washington, D.C.: Albert Shaker Institute.
- *Englund, L. W. (2010). *Evaluating and improving the quality of teacher's language modeling in early childhood classrooms*. University of Nevada, Las Vegas, NV.
- Erning, G. (2003). *Qualitätsentwicklung in Kindergärten. Abschlussbericht der Fortbildung 2001-2003*. Bamberg: Kreisjugendamt Bamberg. Abt. Kindertagesstätten
- Espinosa, L., & Chen, W.-J. (1996). The effect of teacher inservice training on technology and multiage grouping: Year one evaluation of constructing and networking for multi-age learning project. *Journal of Computing in Childhood Education*, 7(1-2), 13-38.
- Espinosa, L. M. (2002). High-quality preschool: Why we need it and what it looks like. *Pre-school Policy Brief, Volume 1*. Retrieved December 11th, 2012, from <http://nieer.org/resources/policybriefs/1.pdf>
- European Union. (2012). Empfehlungen des Rates vom 20. Dezember 2012 zur Validierung nichtformalen und informellen Lernens (2012/C 398/01). Retrieved December 27th, 2013, from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:398:0001:0005:DE:PDF>
- Eurydice. (2014). *Key Data on Early Childhood Education and Care 2014*. Retrieved July 4th, 2014: from http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/166EN.pdf
- Everston, C. M., & Burry, J. (1989). Capturing classroom context: The observation system as lens for assessment. *Journal of Personal Evaluation in Education*, 2(4), 297 -320.
- Evertson, C. M., & Smithey, M. W. (2000). Mentoring effects of protégés' classroom practice: An experimental field study. *The Journal of Educational Research*, 93(5), 294-304.
- Expertengruppe Berufsbegleitende Weiterbildung. (2013). Qualität in der Fort- und Weiterbildung von pädagogischen Fachkräften in Kindertageseinrichtungen. Standards, Indikatoren und Nachweismöglichkeiten für Anbieter. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Kooperationen, Band 2. Retrieved November, 21st, 2013,

- from http://www.weiterbildungsinitiative.de/uploads/media/Koop_02_Expertengruppe-Berufsbegleitende-Weiterbildung.pdf
- Facteau, J. D., Dobbins, G. H., Russell, J. E. A., Ladd, R. T., & Kudisch, J. D. (1995). The influence of general perceptions of the training environment on pre-training motivation and perceived training transfer. *Journal of Management*, 21(11), 1-25.
- Fiene, R. (2002). Improving child care quality through an infant caregiver mentoring project. *Child & Youth Care Forum*, 31(2), 79-87.
- Filcheck, H. A., McNeil, C. B., Greco, L. A., & Bernard, R. S. (2004). Using a whole-class token economy and coaching of teacher skills in a preschool classroom to manage disruptive behavior. *Psychology in the Schools*, 41(3), 351-361.
- Fitzer, S. E. (2009). *The impact of professional development on literacy achievement of kindergarten hispanic english language learners*. Unpublished dissertation, Bowie State University, Bowie.
- Fleishman, E. A., & Hempel, W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, 49(5), 301-312.
- *Flowers, H., Girolametto, L., Weitzman, E., & Greenberg, J. (2007). Promoting early literacy skills: Effects of in-service education for early childhood educators. *Canadian Journal of Speech-Language Pathology and Audiology*, 31(1), 6-18.
- Ford, J. K., & Kraiger, K. (1995). The application of cognitive constructs and principles to the instructional systems model of training: Implications for needs assessment, design, and transfer. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 10) (pp. 1-48). Chichester, UK: Wiley.
- Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology*, 83(2), 218-233.
- Forry, N., Tout, K., Zaslow, M., & Martinez-Beck, I. (2011). *Coaching, consultation and on-site quality improvement: Results from recent research*. Paper presented at the NAEYC 20th National Institute for Early Childhood Professional Development, Providence, RI.
- Fried, L., & Briedigkeit, E. (2008). *Sprachförderkompetenz – Selbst und Teamqualifizierung für Erzieherinnen, Fachberatung und Ausbilder*. Berlin, Düsseldorf, Mannheim: Cornelsen Verlag.
- Fröhlich-Gildhoff, K., Nentwig-Gesemann, I., & Pietsch, S. (2011). *Kompetenzorientierung in der Qualifizierung frühpädagogischer Fachkräfte. Eine Expertise der Weiterbildungsinitiative Frühpädagogische Fachkräfte (WiFF)*. München: Deutsches Jugendinstitut.
- Fröhlich-Gildhoff, K., & Röser, C. (2013). Zertifizierungsinitiative Frühpädagogik Südbaden (ZFS). Ein Modellprojekt zur Anrechnung außerhochschulisch erworbener Kompetenzen auf ein Studium der Frühpädagogik. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Kooperationen, Band 5. Retrieved December 27th, 2013, from http://www.weiterbildungsinitiative.de/uploads/media/Koop_5_ZFS.pdf
- Fuchs-Rechlin, K. (2007). Wie gehts im Job? KiTa-Studie der GEW. Retrieved February 21st, 2013, from <http://www.gew.de/Binaries/Binary35437/GEW-Kitastudie.pdf>
- Fukkink, R. G., & Lont, A. (2007). Does training matter? A meta-analysis and review of caregiver training studies. *Early Childhood Research Quarterly*, 22(3), 294-311.
- Fukkink, R. G., & Tavecchio, L. W. C. (2010). Effects of video interaction guidance on early childhood teachers. *Teaching and Teacher Education: An International Journal of Research and Studies*, 26(8), 1652-1659.
- Fukkink, R. G., Trienekens, N., & Kramer, L. J. C. (2011). Video Feedback in Education and Training: Putting Learning in the Picture. *Educational Psychology Review*, 23(1), 45-63.
- *Gallagher, P. A., Abbott-Shim, M., & VandeWiele, L. (2011). An evaluation of the individualized learning intervention: A mentoring program for early childhood teachers. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(2), 57-74.

- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38(4), 915-945.
- Garvey, W. D., & Griffith, B. C. (1979). Scientific communication as a social system. In W. D. Garvey (Ed.), *Communication: The essence of science* (pp. 148-164). Oxford, UK: Pergamon.
- Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, 27(4), 513-543.
- Gaudine, A. P., & Saks, A. M. (2004). A longitudinal quasi-experiment on the effects of post-training transfer interventions. *Human Resource Development Quarterly*, 15(1), 57-76.
- Gebrande, J. (2011). Der Deutsche Qualifikationsrahmen (DQR) im Lichte fach- und bildungspolitischer Stellungnahmen. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Expertisen, Band 17. Retrieved March 18th, 2012, from http://www.weiterbildungsinitiative.de/uploads/media/WiFF_Expertise_Nr._17_Johanna_Gebrande_Der_Deutsche_Qualifikationsrahmen.pdf
- Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, 18(3), 271-279.
- Gillanders, C., Iruka, I., Ritchie, S., & Cobb, C. T. (2012). Restructuring and aligning early education opportunities for cultural, language, and ethnic minority children. In R. C. Pianta, W. S. Barnett, L. M. Justice & S. M. Sheridan (Eds.), *Handbook of early childhood education* (pp. 111-136). New York, NY: Guilford.
- *Girard, L.-C., Girolametto, L., Weitzman, E., & Greenberg, J. (2011). Training early childhood educators to promote peer interactions: Effects on children's aggressive and prosocial behaviors. *Early Education & Development*, 22(2), 305-323.
- *Girolametto, L., Weitzman, E., & Greenberg, J. (2003). Training day care staff to facilitate children's language. *American Journal of Speech-language Pathology*, 12(3), 299-311.
- *Girolametto, L., Weitzman, E., & Greenberg, J. (2004). The effects of verbal support strategies on small-group peer interaction. *Language, Speech, and Hearing Services in Schools*, 35(3), 254-268.
- Girolametto, L., Weitzman, E., & Greenberg, J. (2012). Facilitating emergent literacy: Efficacy of a model that partners speech-language pathologists and educators. *American Journal of Speech-Language Pathology*, 21(1), 47-63.
- *Girolametto, L., Weitzman, E., Lefebvre, P., & Greenberg, J. (2007). The effects of in-service education to promote emergent literacy in child care centers: A feasibility study. *Language, Speech, and Hearing Services in Schools*, 38(1), 72-83.
- Gist, M. E. (1989). The influence of training method on self-efficacy and idea generation among managers. *Personnel Psychology*, 42(4), 787-805.
- Gist, M. E., & Stevens, C. K. (1998). Effects of practice conditions and supplemental training method on cognitive learning and interpersonal skill generalization. *Organizational Behavior & Human Decision Processes*, 75(2), 142-169.
- Gist, M. E., Stevens, C. K., & Bavetta, A. G. (1991). Effects of self-efficacy and post-training intervention on the acquisition and maintenance of complex interpersonal skills. *Personnel Psychology*, 44(4), 837-861.
- Gnefkow, T. (2008). *Lerntransfer in der betrieblichen Weiterbildung. Determinanten für den Erfolg externer betrieblicher Weiterbildungen im Lern- und Funktionsfeld aus Teilnehmerperspektive*. Universität Bielefeld.
- Goldstein, I., & Ford, J. K. (2002). *Training in organizations* (4th ed.). Belmont, CA: Wadsworth.
- *Grace, C., Bordelon, D., Cooper, P., Kazelskis, R., Reeves, C., & Thames, D. G. (2008). Impact of professional development on the literacy environments of preschool classrooms. *Journal of Research in Childhood Education*, 23(1), 52-75.
- *Graul, S. K., & Zeece, P. D. (1990). Effects of play training of adults on the cognitive and play behavior of preschool children. *Early Child Development and Care*, 57(1), 15-22.

- Greve, W., & Wentura, D. (1997). *Wissenschaftliche Beobachtung: Eine Einführung*. Weinheim: PVU/Beltz.
- Grimm, H., & Weinert, S. (2008). Sprachentwicklung. In R. Oerter & L. Montada (Eds.), *Entwicklungspsychologie* (pp. 517-550). Weinheim, Basel, Berlin: Beltz.
- Gupta, S. S., & Daniels, J. (2012). Coaching and professional development in early childhood classrooms: Current practices and recommendations for the future. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 15(2), 206-220.
- Hall, J. A., Rosenthal, R., Tickle-Degnen, L., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 17-28). New York, NY: Russell Sage.
- Hamre, B. K., & Hatfield, B. E. (2012). Moving evidenced-based professional development into the field: Recommendations for policy and research. In C. Howes, B. Hamre, & R. Pianta (Eds.), *Effective early childhood professional development. Improving teacher practice and child outcomes* (pp.213-228). Baltimore, MD: Brookes.
- Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., et al. (2010). Implementation fidelity of my teaching partner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, 25(3), 329-347.
- Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., & Downer, J. T. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, 49(1), 88-123.
- Harbord, R. M., Harris, R. J., & Sterne, J. A. C. (2009). Updated test for small-study effects in meta-analyses. *The Stata Journal*, 9(2), 197-210.
- Hartmann, E., Rvachew, S., & Grawburg, M. (2008). Phonological awareness in preschoolers with spoken language impairments: Toward a better understanding of causal relationships and effective intervention. A constructive comment on Rvachew and Grawburg's (2006) study. *Journal of Speech, Language, and Hearing*, 51(5), 1215-1218.
- Harvard Family Research Project. (2006). Pathways from workforce development to child outcomes. *The Evaluation Exchange*, 11(4), 2-4.
- Harvey, E., Fischer, C., Weineth, J. L., Hurwitz, S. D., & Sayer, A. G. (2013). Predictors of discrepancies between informants' ratings of preschool-aged children's behavior: An examination of ethnicity, child characteristics, and family functioning. *Early Childhood Research Quarterly*, 28(4), 668-682.
- Haskell, M. (1994). *Using training as a means to improve the level of quality in child care facilities*. Unpublished dissertation, Nova Southeastern University, Fort Lauderdale.
- Hatfield, B., Hamre, B., LaCosale-Crouch, J., Pianta, R., Downer, J., Burchinal, M., et al. (2012). *Teacher characteristics influence responsiveness to a course and a consultancy focused on effective teacher-child interactions*. Paper presented at the Society for Research on Educational Effectiveness Conference, Washington, D.C.
- Heckman, J. (2008). Schools, skills, and synapses. *Economic Inquiry*, 40(3), 289-324.
- Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, 2(3), 167-171.
- Hedges, L. V. (2009). Effect sizes in studies with nested designs. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd Edition*. (pp. 337-356) New York : The Russell Sage Foundation.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Helburn, S. (1995). *Cost, quality, and child outcomes in child care centers. Technical report*. Denver, CO: Department of Economics, Center for Research in Economic and Social Policy, University of Colorado.
- *Helmer, J., Bartlett, C., Wolgehmut, J. R., & Lea, T. (2011). Coaching (and) commitment: Linking ongoing professional development, quality teaching and student outcomes. *Professional Development in Education*, 37(2), 197-211.

- Hemmeter, M. L., Snyder, P., Kinder, K., & Artman, K. (2011). Impact of performance feedback delivered via electronic mail on preschool teachers' use of descriptive praise. *Early Childhood Research Quarterly*, 26(1), 96-109.
- Hendrickson, J. M., Gardner, N., Kaiser, A., & Riley, A. (1993). Evaluation of a social interaction coaching program in an Integrated day-care setting. *Journal of Applied Behavior Analysis*, 26(2), 213-225.
- Henk, J. K., Morrison, J. W., Thornburg, K. R., & Raya-Carlton, P. (2007). The efficacy of HeadsUp! Reading in Missouri on teachers' knowledge of emergent literacy: A satellite-based literacy development training course. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 10(1), 20-35.
- Higgins, J. P. T., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic review of interventions* (pp. 187-241). Chichester, UK: Wiley-Blackwell.
- Higgins, J. P. T., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. Retrieved 8th April, 2013, from www.cochrane-handbook.org.
- High/Scope Educational Research Foundation. (1989). *Primary Implementation Profile (PIP) administration manual*. Ypsilanti, MI: The High/Scope Press.
- Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.
- Hindman, A. H., & Wasik, B. A. (2012). Unpacking an effective language and literacy coaching intervention in Head Start. *The Elementary School Journal*, 113(1), 131-154.
- Hoff, E., Core, C., & Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-months-olds. *Journal of Child Language* 35(4), 903-916.
- Honig, B. (1996). *Teaching our children to read: The role of skills in a comprehensive reading program*. Thousand Oaks, CA: Corwin Press.
- *Howe, N., Jacobs, E., Vukelich, G., & Recchia, H. (2011). In-service professional development and constructivist curriculum: Effects on quality of child care, teacher beliefs, and interactions. *Alberta Journal of Educational Research*, 57(7), 353-378.
- Howes, C. (1988). Relations between early child care and schooling. *Developmental Psychology*, 24(1), 53-57.
- Howes, C. (1997). Children's experiences in center-based child care as a function of teacher background and adult: Child ratio. *Merrill-Palmer Quarterly*, 43(3), 404-425.
- Howes, C., Burchinal, M., Pianta, R. C., Bryant, D., Early, D. M., & Clifford, R. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, 23(1), 27-50.
- Howes, C., Whitebook, M. & Phillips, D. (1992). Teacher characteristics and effective teaching in child care: Findings from the national child care staffing study. *Child and Youth Care Forum*. 21(6), 399-414.
- Hox, J. (2010). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- *Hsieh, W.-Y. (2005). *Teaching strategies that support emergent literacy: Effects of a collaborative training model*. Unpublished dissertation, University of Illinois, Urbana-Champaign.
- Hsieh, W.-Y., Hemmeter, M. L., McCollum, J., & Ostrosky, M. M. (2009). Using coaching to increase preschool teachers' use of emergent literacy teaching strategies. *Early Childhood Research Quarterly*, 24(3), 229-247.
- Huczynski, A. A., & Lewis, J. W. (1980). An empirical study into the learning transfer process in management training. *Journal of Management Studies*, 17(2), 227-240.
- Huedo-Medina, T., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q-statistic or I² index? *CHIP Documents*. Paper 19. Retrieved May 17th, 2013, from http://digitalcommons.uconn.edu/cgi/viewcontent.cgi?article=1019&context=chip_docs
- Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis. Correcting error and bias in search findings*. Thousand Oakes, CA: Sage Publications.

- Ihle, W., & Esser, G. (2002). Epidemiologie psychischer Störungen im Kindes- und Jugendalter: Prävalenz, Verlauf, Komorbidität und Geschlechterunterschiede. *Psychologische Rundschau*, 53(4), 159-169.
- Incredible Years. (n.d.). *The Incredible Years Training Series: The Incredible Years parent, teacher and child programs. Overview of program details*. Seattle, WA: Incredible Years.
- Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes and efficacy. *Education Policy Analysis Archives*, 13(1), 1-28.
- Institute of Medicine & National Research Council of the National Academies (2012). *The early childhood care and education workforce. Challenges and opportunities. A workshop report*. Washington, D.C.: The National Academies Press.
- Isner, T., Tout, K., Zaslow, M., Soli, M., Quinn, K., Rothenberg, L., et al. (2011). *Coaching in early care and education programs and Quality Rating and Improvement Systems (QRIS): Identifying promising features*. Washington, D.C.: Child Trends.
- Iutovich, J., Fiene, R., Johnson, J., Koppel, R., & Langan, F. (1997). *Investing in our children's future: The path to quality child care through the Pennsylvania Child Care/Early Childhood Development Training System*. Erie, PA: Keystone University Research Corporation.
- Ivancic, K., & Hesketh, B. (2000). Learning from errors in a driving simulation: Effects on driving skill and self confidence. *Ergonomics*, 43(12), 1966–1984.
- Jackson, B., Larzelere, R., St. Clair, L., Corr, M., Fichter, C., & Egertson, H. (2006). The impact of "HeadsUp! Reading" on early childhood educators' literacy practices and preschool children's literacy skills. *Early Childhood Research Quarterly*, 21(2), 213-226.
- *Japel, C. (2009). *How can we enhance child care quality?* Paper presented at the ECERS International Network Meeting, Porto.
- Jo Rodriguez, B., Loman, S. L., & Horner, R. H. (2009). A preliminary analysis of the effects of coaching feedback on teacher implementation fidelity of first step to success. *Behavior Analysis in Practice*, 2(2), 11-21.
- Joyce, B., & Showers, B. (2002). *Student achievement through staff development*. Alexandria, VA: ASCD.
- Joyce, B. R., & Showers, B. (1981). Transfer of training: The contribution of "coaching". *Journal of Education*, 163(2), 163-172.
- Jude, N. (2008). *Zur Struktur von Sprachkompetenz*. Unpublished dissertation, Johann Wolfgang Goethe Universität, Frankfurt am Main.
- Jungmann, T., Koch, K., & Etzien, M. (2013). Effektivität alltagsintegrierte Sprachförderung bei ein- und zwei- bzw. mehrsprachig aufwachsenden Vorschulkindern. *Frühe Bildung*, 2(3), 110-121.
- Jüni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: Empirical study. *International Journal of Epidemiology*, 31(1), 115-123.
- Justice, L. M., & Ezell, H. K. (2001). Word and print awareness in 4-year-old children. *Child Language Teaching and Therapy*, 17(3), 207-225.
- Justice, L. M., Mashburn, A. J., Hamre, B. K., & Pianta, R. C. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, 23(1), 51-68.
- Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74(4), 657–690.
- Karoly, L. A., Ghosh-Dastidar, B., Zellman, G. L., Perlman, M., & Fernyhough, L. (2008). Prepared to learn: The nature and quality of early care and education for preschool-age children in California. Retrieved February 17th, 2011, from http://www.rand.org/content/dam/rand/pubs/technical_reports/2008/RAND_TR539.pdf
- Kelley, P., & Camilli, G. (2007). The impact of teacher education on outcomes in center-based early childhood education programs: A meta-analysis. NIEER Working Paper.

- Retrieved January 15th, 2011, from
<http://nieer.org/resources/research/TeacherEd.pdf>
- Klein, L., & Gomby, D. S. (2008). *A synthesis of federally funded studies on school readiness: What are we learning about professional development? Prepared for "A Working Meeting on Recent School Readiness Research: Guiding the Synthesis of Early Childhood Research,"* Washington, D.C.: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation (ASPE) and the Administration for Children and Families, Office of Planning Research and Evaluation (OPRE).
- Kontoghiorghes, C. (2001). Factors affecting training effectiveness in the context of the introduction of new technology – A US case study. *International Journal of Training and Development*, 5(4), 248–260.
- Kontos, S., Howes, C., & Galinsky, E. (1996). Does training make a difference to quality in family child care? *Early Childhood Research Quarterly*, 11(4), 427-445.
- *Kopacsi, R. & Hochwald, E. (1998). Restructuring kindergarten in an urban district: The case of Newark, New Jersey. Retrieved June 25th, 2012, from
<http://files.eric.ed.gov/fulltext/ED422104.pdf>
- Krause, M. & Kofler, A. (2011). *Evaluation – Teilstudie im Raum Regensburg. Endbericht Sprachberatung*. München: IFP Staatsinstitut für Frühpädagogik.
- *Krause, M., Kofler, A., & Hofbauer, C. (2011). Lassen sich Effekte von Sprachberatung messen? *Fachkongress "Sprachliche Bildung von Anfang an - Strategien, Konzepte und Erfahrungen"* Retrieved January 2nd, 2012, from
http://www.ifp.bayern.de/imperia/md/content/stmas/ifp/fachkongress/poster_regensburger_studie.pdf
- Krijger, N. L., & Pol, S. M. (1995). Changing leadership style: A training model for lasting effects. In M. Mulder, W. J. Nijhof & R. O. Brinkerhoff (Eds.), *Corporate training for effective performance: Evaluation in education and human services* (pp. 111-133). Norwell, MA: Kluwer Academic.
- Kruse, Elke (2012): Anrechnung beruflicher Kompetenzen von Erzieherinnen und Erziehern auf ein Hochschulstudium. Konzepte – Erfahrungen – Prozesse. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Expertisen, Band 28. München. Retrieved May 17th, 2013, from
http://www.weiterbildungsinitiative.de/uploads/media/Expertise_Kruse.pdf
- Kuger, S., Sechtig, J., & Anders, Y. (2012). Kompensatorische (Sprach-)Förderung: Was lässt sich aus US-amerikanischen Projekten lernen? *Frühe Bildung*, 1(4), 181-193.
- Ladd, G. W., Price, J. M., & Hart, C. H. (1988). Predicting preschoolers' peer status from their playground behaviors. *Child Development*, 59(4), 986-992.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Landry, S. (2005). Effective early childhood programs: Turning knowledge into action. Retrieved January 2nd, 2014, from
http://www.childrenslearninginstitute.org/library/publications/documents/Effective-Early_Childhood-Programs.pdf
- *Landry, S. H., Anthony, J. L., Swank, P. R., & Monseque-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology*, 101(2), 448-465.
- *Landry, S. H., Swank, P. R., Anthony, J. L., & Assel, M. A. (2011). An experimental study evaluation professional development activities within a state funded pre-kindergarten program. *Reading and Writing*, 24(8), 971-1010.
- Landry, S. H., Swank, P. R., Smith, K. E., Assel, M. A., & Gunnewig, S. B. (2006). Enhancing early literacy skills for preschool children: Bringing a professional development model to scale. *Journal of Learning Disabilities*, 39(4), 306-324.
- Laughlin, L. (2013). Who's minding the kids? Child care arrangements: Spring 2011. Household Economic Studies. Retrieved June 2nd, 2013, from
<http://www.census.gov/prod/2013pubs/p70-135.pdf>

- Lavigne, J. V., Gibbson, R. D., Christoffel, K. K., Arend, R., Rosenbaum, D., Binns, H., et al. (1996). Prevalence rates and correlates of psychiatric disorders among preschool children. *Journal of the American Child and Adolescent Psychiatry*, 35(2), 203-214.
- Lee, C. D., & Kahnweiler, W. M. (2000). The effect of a mastery learning technique on the performance of a transfer of training task. *Performance Improvement Quarterly*, 13(3), 125-139.
- *Lee, Y. (2010). Blended teacher support for promoting open-ended questioning in pre-k science activities. Unpublished dissertation. University of Virginia, Charlottesville.
- Lim, D. H., & Morris, M. L. (2006). Influence of trainee characteristics, instructional satisfaction, and organizational climate on perceived learning and training transfer. *Resource Development Quarterly*, 17(1), 85-115.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181-1209.
- *Lloyd, C. M., & Millenky, M. (2011). *Teacher training, classroom consultation, and child outcomes in the Foundations of Learning Project*. Paper presented at the Society for Research on Educational Effectiveness Conference, Washington, D.C.
- Lloyd, C. M., & Modlin, E. L. (2012). *Coaching as a key component in teachers' professional development. Improving classroom practices in Head Start settings. OPRE Report 2012-4*. Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- LoCasale-Crouch, J., Kraft-Sayre, M., Pianta, R., Hamre, B. K., Downer, J. T., Leach, A., et al. (2011). Implementing an early childhood professional development course across 10 sites and 15 sections: Lessons learned. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 275-292.
- Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90(1), 125-152.
- Lonigan, C. J., Clancy-Menchetti, J., Phillips, B. M., McDowell, K., & Farver, J. M. (2005). *Literacy Express: A preschool curriculum*. Tallahassee, FL: Literacy Express.
- *Lonigan, C. J., Farver, J. M., Phillips, B. M., & Clancy-Menchetti, J. (2011). Promoting the development of preschool children's emergent literacy skills: A randomized evaluation of a literacy-focused curriculum and two professional development models. *Reading and Writing: An Interdisciplinary Journal*, 24(3), 305-337.
- Manningham, S. (2008). *Qualité de l'environnement éducatif dans les services de garde préscolaires au Québec : Rôle des caractéristiques des éducatrices et une intervention visant à augmenter ses compétences*. Unpublished dissertation, Université de Montréal, Montréal.
- Martinez-Beck, I. (2011). Introduction: Why strengthening the measurement of quality in early childhood settings has taken new importance. In M. Zaslow, I. Martinez-Beck, K. Tout & T. Halle (Eds.), *Quality measurement in early childhood settings* (pp. xviii-xxiv). Baltimore, MD: Brookes.
- Martinez-Beck, I., & Zaslow, M. (2006). Introduction: The context for critical issues in early childhood professional development. In M. Zaslow & I. Martinez-Beck (Eds.), *Critical issues in early childhood professional development* (pp. 1-16). Baltimore, MD: Brookes.
- *Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, 14(4), 179-196.
- Mashburn, A. J., Pianta, R., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development*, 79(3), 732-749.
- Mason, E. B. (2010). *Educating minority boys: Examining the differences in children's classroom experiences as a result of teacher-child relationships and a professional development*. Unpublished master thesis, University of North Carolina at Chapel Hill, Chapel Hill.

- Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal*, 35(4), 828-847.
- Maxwell, K. L., Feild, C. C., & Clifford, R. M. (2006). Defining and measuring professional development in early childhood research. In M. Z. I. Martinez-Beck (Ed.), *Critical issues in early childhood professional development* (pp. 21-48). Baltimore, MD: Brookes.
- McCartney, K. (1984). The effects of quality of day care environment upon children's language development. *Developmental Psychology*, 20(2), 244-260.
- *McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., et al. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning *Journal of Learning Disabilities*, 35(1), 69-86.
- McEvoy, M. A., Odom, S. L., & McConnell, S. R. (1992). Peer social competence intervention for young children with disabilities. In M. A. McEvoy, S. L. Odom & S. R. McConnell (Eds.), *Social competences of young children with disabilities: Issues and strategies for intervention* (pp. 113-133). Baltimore, MD: Brookes.
- *McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *The Journal of Educational Research*, 93(2), 67-74.
- *McNerney, S., Nielsen, D. C., & Clay, P. (2006). Supporting literacy in preschool: Using a teacher-observation tool to guide professional development. *Journal of Early Childhood Teacher Education*, 27(1), 19-34.
- Miller, J. A., & Bogatova, T. (2009). Quality improvements in the early care and education workforce: Outcomes and impact of the T.E.A.C.H. Early Childhood Project. *Evaluation and Program Planning*, 32(3), 257-277.
- Ministère de la Famille et des Aînés (MFA). (2007). Accueillir la petite enfance. Le programme éducatif des service de garde du Québec. Retrieved June 2, 2013, from http://www.mfa.gouv.qc.ca/fr/publication/Documents/programme_educatif.pdf
- Moher, D., Fortin, P., Jadad, A. R., Jüni, P., Klassen, T. P., Le Lorier, J., Liberati, A., Linde, K., & Penna, A. (1996). Completeness of reporting of trials published in languages other than English: Implications for conduct and reporting of systematic reviews. *The Lancet*, 347(8998), 363-366.
- Moher D., Pham, B., Lawson, M. L., Klassen, T. P., Schulz, K. F., Berlin, J. A., Jadad, A. R., & Liberati, A. (2000). What contributions do languages other than English make on the results of meta-analyses? *Journal of Clinical Epidemiology*, 53(9), 964-972.
- Moher D., Pham, B., Lawson, M. L., & Klassen, T. P. (2003). The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technology Assessment*, 7(41), 1-90.
- *Mohler, G. M., Yun, K. A., Carter, A., & Kasak, D. (2009). The effect of curriculum, coaching, and professional development on prekindergarten children's literacy achievement. *Journal of Early Childhood Teacher Education*, 30(1), 49-68.
- Mol, S. E. & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, 137(2), 267-296.
- Morin, L., & Latham, G. P. (2000). The effect of mental practice and goal setting as a transfer of training intervention on supervisors' self-efficacy and communication skills: An exploratory study. *Applied Psychology: An International Review*, 49(3), 566-578.
- *Morris, P., Raver, C. C., Millenky, M., Jones, S., Llyod, C. M. (2010). Making preschool more productive. How classroom management training can help teachers. Retrieved, January 15th, 2011, from <http://files.eric.ed.gov/fulltext/ED514651.pdf>
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1), 105-125.
- Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: The Brookings Institute.
- Mulrow, C. D. (1994). Rational for systematic reivews. *British Medical Journal*, 309(6954), 597-599.

- National Association for the Education of Young Children (NAEYC), & National Association of Child Care Resources & Referral Agencies (NACCRRA). (2011). Early childhood education professional development: Training and technical assistance glossary. Retrieved June 16th, 2013, from http://www.naeyc.org/GlossaryTraining_TA.pdf
- National Association for the Education of Young Children (NAEYC), & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE). (2002). Early learning standards: Creating the condition for success. Retrieved September 22nd, 2012, from http://www.naeyc.org/files/naeyc/file/positions/position_statement.pdf
- National Association for the Education of Young Children (NAEYC), & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE). (2009). Where we stand on early learning standards. Retrieved September 22nd, 2012, from <http://www.naeyc.org/files/naeyc/file/positions/earlyLearningStandards.pdf>
- National Evaluation of Sure Start NESS Team. (2010). The impact of Sure Start Local Programmes on five year olds and their families. Retrieved March 16th, 2013, from <http://www.ness.bbk.ac.uk/impact/documents/RR067.pdf>
- National Institute for Literacy (2008). Developing early literacy. Report from the National Reading Panel. Retrieved August 6th, 2012 from <http://lincs.ed.gov/publications/pdf/NELPReport09.pdf>
- National Institute of Child Health and Human Development, E. C. C. R. N. [NICHD] (2000a). The relation of child care to cognitive and language development. *Child Development*, 74(4), 960–980.
- National Institute of Child Health and Human Development, E. C. C. R. N. (2000b). Characteristics and quality of child care for toddlers and preschoolers. *Applied Developmental Science*, 4(3), 116-135.
- National Institute of Child Health and Human Development, E. C. C. R. N. (2002). CHILD CARE STRUCTURE àPROCESS àOUTCOME: Direct and indirect effects of child-care quality on young children's development. *Psychological Science*, 13(3), 199-206.
- National Institute of Child Health and Human Development, E. C. C. R. N. (2003). Does quality of child care affect child outcomes at age 4 1/2? *Developmental Psychology* 39(3), 451-469.
- National Institute of Child Health and Human Development, E. C. C. R. N., & Duncan, G. (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, 74(5), 1454–1475.
- National Professional Development Center on Inclusion. (2008). *What do we mean by professional development in the early childhood field?* Chapel Hill, NC: The University of North Carolina, FPG Child Development Institute.
- National Reading Panel (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, D.C.: US Government Printing Office.
- National Research Council. (2004a). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, D.C.: National Academies Press.
- National Research Council. (2004b). Implementing randomized field trials in education: Report of a workshop. Retrieved June 4th, 2013, from <http://mlemons.perl.hs.iastate.edu/sites/default/files/uploads/projects/carver/randomtrials.pdf>
- National Research Council. (2005). Advancing scientific research in education. Retrieved June 4th, 2013, from http://biblioteca.ucv.cl/site/colecciones/manuales_u/11112.pdf
- Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention and Treatment*, 6, 1-34.
- *Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal*, 46(2), 532-566.

- Neuman, S. B., & Dickinson, D. (2001). *Handbook of early literacy research*. Volume 1. New York, NY: Guilford.
- *Neuman, S. B., & Wright, T. S. (2010). Promoting language and literacy development for early childhood educators: A mixed-methods study of coursework and coaching. Retrieved April 14th, 2011, from <http://www-personal.umich.edu/~sbneuman/pdf/coachingInEarlyChildhood.pdf>
- Oberhuemer, P. (2012). Fort- und Weiterbildung frühpädagogischer Fachkräfte im europäischen Vergleich. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WIFF Studien, Band 17. Retrieved April, 19th, 2013, from http://www.weiterbildungsinitiative.de/uploads/media/Studie_Oberhuemer.pdf
- Oberhuemer, P. (2012). *Fort- und Weiterbildung frühpädagogischer Fachkräfte im europäischen Vergleich. WIFF Studien. Band, 17*. München: Deutsches Jugendinstitut.
- Oberhuemer, P. (2013). Continuing professional development and the early years workforce. *Early Years: An International Research Journal*, 33(2), 103-105.
- *O'Connor, R. E. (1999). Teachers learning Ladders to Literacy. *Learning Disabilities Research and Practice*, 14(4), 203-214.
- OECD. (2012). *Starting Strong III: A quality toolbox for early childhood education and care*. Paris: OECD Publishing.
- OECD. (n.d.). International comparison: Professional development. Retrieved February 17, 2013, from <http://www.oecd.org/edu/school/49322277.pdf>
- Olson, R. K., Keenan, J. M., Byrne, B., Samuelsson, S., Coventry, W. L., Corley, R., et al. (2007). Genetic and environmental influence on vocabulary and reading development. *Scientific Studies of Reading*, 20(1-2), 51-75.
- *Onchwari, G., & Keengwe, J. (2010). Teacher mentoring and early literacy learning: A case study of a mentor-coach initiative. *Early Childhood Education Journal* 37(4), 311-317.
- *Palsha, S. A., & Wesley, P. W. (1998). Improving quality in early childhood environments through on-site consultation. *Topics in Early Childhood Special Education*, 18(4), 243-153.
- Peisner-Feinberg, E., & Burchinal, M. (1997). Relations between preschool children's child care experiences and concurrent development: The Cost, Quality and Outcomes Study. *Merrill-Palmer Quarterly*, 43(3), 451-477.
- Peisner-Feinberg, E., Burchinal, M., Clifford, R., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development* 72(5), 1534-1553.
- Peisner-Feinberg, E., Burchinal, M., Clifford, R., Yazejian, N., Culkin, M., Zelazo, J., et al. (1999). *The children of the Cost, Quality, and Outcomes Study go to school*. Chapel Hill, NC: University of North Carolina at Chapel Hill, FPG Child Development Institute.
- Penuel, W. R., J., F. B., Yamaguchi, R., & Gallagher, L. P. (2007). What make professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, 44(4), 921-958.
- Peucker, C., Gragert, N., Pluto, L., & Seckinger, M. (2010). *Kindertagesbetreuung unter der Lupe. Befunde zu Ansprüchen an eine Förderung von Kindern. Reihe: DJI-Fachforum Bildung und Erziehung, Band 9*. München: Deutsches Jugendinstitut.
- Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly*, 12(3), 281-303.
- Pianta, R. C., Barnett, W. S., Burchinal, M. P., & Thornburg, K. R. (2009). The effects of pre-school education: What we know, how public policy is or is not aligned with the evidence base, and what we don't know. *Psychological Science in the Public Interest*, 10(2), 49-88.
- Pianta, R. C., Burchinal, M., Jamil, F. M., Sabol, T., Grimm, K., Hamre, B. K., et al. (2014). A cross-lag analysis of longitudinal associations between preschool teachers' instructional support identification skills and observed behavior. *Early Childhood Research Quarterly*, 29(2), 144-154.

- Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, 23(4), 431-451.
- Piasta, S. B., & Wagner, R. K. (2010). Developing early literacy skills: A meta-analysis of alphabet learning and instruction. *Reading Research Quarterly*, 45(1), 8-38.
- Piasta, S. B., Justice, L. M., Cabell, S. Q., Wiggins, A. K., Turnbull, K. P., & Curenton, S. M. (2012). Impact of professional development on preschool teachers' conversational responsivity and children's linguistic productivity and complexity. *Early Childhood Research Quarterly*, 27(3), 387-400.
- *Podhajski, B., & Nathan, J. (2005). Promoting early literacy through professional development for childcare providers. *Early Education and Development*, 16(1), 23-41.
- Powell, D., Diamond, K., & Burchinal, M. (2009). *Effects of a professional development intervention on teaching practices and children's language and literacy outcomes*. Chapel Hill, NC: University of North Carolina at Chapel Hill.
- Powell, D. R., & Diamond, K. E. (2011). Improving the outcomes of coaching-based professional development. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 295-307). New York, NY: Guilford.
- *Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology*, 102(2), 299-312.
- Preschool Curriculum Evaluation Research Consortium. (2008). Effects of preschool curriculum programs on school readiness. Retrieved June 13th, 2012, from http://ies.ed.gov/ncer/pubs/20082009/pdf/20082009_1.pdf
- Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work. *Journal of Applied Psychology*, 68(1), 78-87.
- Rank, B., & Wakenhut, R. (1998). Ein Bedingungsmodell des Praxistransfers. In B. Rank & R. Wakenhut (Eds.), *Sicherung des Praxistransfers im Führungskräfte training* (pp. 11-29). München: Hampp.
- *Rasberry, J. E. (2004). Early childhood literacy enrichment training of teachers and of three and four year olds in a preschool setting. Unpublished dissertation. Mississippi State University, Mississippi.
- Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN Version 2.02. Centre for Multilevel Modelling, University of Bristol.
- Raver, C., & Knitze, J. (2002). *Ready to enter: What research tells policymakers about strategies to promote social and emotional school readiness among 3 and 4 years old*. New York, NY: National Center for Children in Poverty. Columbia University.
- *Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Champion, K. M., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, 23(1), 10-26.
- Reeves, B. C., Deeks, J. J., Higgins, J. P. T., & Wells, G. A. (2008). Including non-randomized studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 391-432). Chichester, UK: Wiley-Blackwell.
- Reischmann, J. (2004). Kompetenz lehren? Kompetenz- und Performanz-Orientierung in der Andragogik zwischen Didaktik und Organisationsentwicklung. Retrieved June 30th, 2009, from <http://www.uni-bamberg.de/fileadmin/andragogik/08/andragogik/aktuelles/Reischmann-Kompetenz.pdf>
- *Rhodes, S., & Hennessy, E. (2000). The effects of specialized training on caregivers and children in early-years settings: An evaluation of the foundation course in playgroup practice. *Early Childhood Research Quarterly*, 15(4), 559-576.
- Richman-Hirsch, W. L. (2001). Posttraining interventions to enhance transfer: The moderating effects of work environments. *Human Resource Development Quarterly*, 12(2), 105-120.
- Riley, J. (1996). *The teaching of reading*. London, UK: Paul Chapman Publishing.

- Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2001). *Emerging Academic Snapshot. Unpublished measure*. Los Angeles, CA: University of California.
- Robert Bosch Stiftung. (2008). Frühpädagogik Studieren - ein Orientierungsrahmen für Hochschulen. Retrieved December 19th, 2012, from http://www.bosch-stiftung.de/content/language2/downloads/PiK_orientierungsrahmen_druckversion.pdf
- Robert Bosch Stiftung. (2011). Qualifikationsprofile in Arbeitsfeldern der Pädagogik der Kindheit. Ausbildungswege im Überblick. Retrieved December 16th, 2012, from http://www.bosch-stiftung.de/content/language2/downloads/RobertBoschStiftung_Studie_PiK_2011_Ausbildungswege.pdf
- Robertson, I., & Downs, S. (1979). Learning and the prediction of performance: Development of trainability testing in the United Kingdom. *Human Resource Development Quarterly*, 64(1), 42-50.
- Rodriguez, C. M., & Gregory, S. (2005). Qualitative study of transfer of training of student employees in a service industry. *Journal of Hospitality & Tourism Research*, 29(1), 42-66.
- *Roehrig, G. H., Dubosarsky, M., Mason, A., Carlson, S., & Murphy, B. (2011). We look more, listen more, notice more: Impact of sustained professional development on Head Start teachers' inquiry-based and culturally-relevant science teaching practices. *Journal of Science Education and Technology*, 20(5), 566-578.
- Rosenthal, M. L. C. (1994). The fugitive literature. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85-94). New York, NY: Russell Sage.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-642.
- *Rosof, H. B. (2006). *The effects of an early childhood professional development program on preschool teachers' knowledge, skills, and confidence of phonological awareness and the impact on student achievement on phonological awareness*. Unpublished dissertation, University of Houston, Houston.
- Roßbach, H.-G. (2005). Effekte qualitativer guter Betreuung, Bildung und Erziehung im frühen Kindesalter auf Kinder und ihre Familien. In S. Z. K.-u. Jugendbericht (Ed.), *Band 1: Bildung, Betreuung und Erziehung von Kindern unter sechs Jahren* (pp. 55-174). München: DJI Verlag.
- Roßbach, H.-G., Kluczniok, K., & Isenmann, D. (2008). Erfahrungen aus internationalen Längsschnittuntersuchungen. In H.-G. Roßbach & S. Weinert (Eds.), *Kindliche Kompetenzen im Elementarbereich: Förderbarkeit, Bedeutung und Messung* (pp. 7-88). Berlin: Bundesministerium für Bildung und Forschung.
- Rouiller, J. Z., & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly*, 4(4), 377-390.
- *Rudd, L. (2003). *The effect of joint attention training for child care providers on the language acquisition of young children*. Unpublished dissertation, Baylor University Waco.
- Rudd, L. C., Lambert, M. C., Satterwhite, M., & Smith, C. H. (2009). Professional development + coaching = enhanced teaching: Increasing usage of math mediated language in preschool classrooms. *Early Childhood Education Journal*, 37(1), 63-39.
- Rush, D. D., & Shelden, M. L. (2011). *The early childhood coaching handbook*. Baltimore, MD: Brookes.
- Rush, D. D., Shelden, M. L., & Hanft, B. E. (2003). Coaching families and colleagues: A process for collaboration in natural settings. *Infants and Young Children*, 16(1), 33-47.
- Rynes, S., & Rosen, B. (1995). A field survey of factors affecting the adoption and perceived success of diversity training. *Personnel Psychology*, 48(4), 247-270.
- Saks, A. M. (1995). Longitudinal field investigation of the moderating and mediating effects of self-efficacy on the relationship between training and newcomer adjustment. *Journal of Applied Psychology*, 80(2), 211-225.
- Sammons, P. (2010). Does pre-school make a difference? Identifying the impact of pre-school on children's cognitive and social behavioral development at different ages. In

- K. Sylva, E. Melhuish, P. Sammons, I. Siraj-Blatchford & B. Taggart (Eds.), *Early childhood matters: Evidence from the Effective Pre-school and Primary Education Project* (pp. 92-113). Oxford, UK: Routledge.
- Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2002). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 8b – Measuring the impact of pre-school on children's social/behavioural development over the pre-school period*. London, UK: DfES / Institute of Education, University of London.
- Samuelsson, S., Olson, R., Wadsworth, S., Corley, R., Defries, J. C., Willcutt, E., et al. (2007). Genetic and environmental influences on prereading skills and early reading and spelling development in the United States, Australia, and Scandinavia. *Reading and Writing*, 20(1-2), 51-75.
- *Schröder, L., Keller, H., Starch, A., & Dintsioudi, A. (2011). *Kultursensitive Sprachförderung im Kita-Alltag - Eine Interventionsstudie*. Paper presented at the 20. Fachgruppentag Entwicklungspsychologie der DGPs, Erfurt.
- Schröder, L., Keller, H., Starch, A., Dintsioudi, A., & List, M. (2012, 12-14th March 2012). *Sprachkultur in der Kita – ein alltagsbasierter Sprachförderansatz*. Paper presented at the 23rd DGFE Congress, Osnabrück.
- Schünemann, H. J., Oxman, A. D., Gunn, V. E., Higgins, J. H. T., Deeks, J., Glasziou, P., et al. (2008). Interpreting results and drawing conclusions. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 359-387). Chichester, UK: Wiley-Blackwell.
- Schweinhart, L. J. (2004). *The High/Scope Perry Preschool Study through age 40: Summary, conclusions, and frequently asked questions*. Ypsilanti, MI: High/Scope Press.
- Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits. The High/Scope Perry Preschool Study through age 27*. Ypsilanti, MI: High Scope Press.
- Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects. The High Scope Perry Preschool Study through age 40*. Ypsilanti, MI: High Scope Press.
- Scott-Little, C., Lesko, J., Martella, J., & Milburn, P. (2007). Early learning standards: Results from a national survey to document trends in state-level policies and practice [Electronic Version]. *Early Childhood Research and Practice*, 9(1). Retrieved 2013, May 17th from <http://ecrp.uiuc.edu/v9n1/little.html>.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454-499.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Retrieved December 11th, 2012, from <https://depts.washington.edu/methods/readings/Shadish.pdf>
- Sheridan, S. (2001). Quality evaluation and quality enhancement in preschool a model of competence development. *Early Child Development and Care*, 166(1), 7-27.
- *Sheridan, S. M., Clarke, B. L., Knoche, L. L., & Edwards, C. P. (2006). The effects of conjoint behavioral consultation in early childhood settings. *Early Education and Development*, 17(4), 593-617.
- Sheridan, S. M., Edwards, C. P., Marcin, C. A., & Knoche, L. L. (2009). Professional development in early childhood programs: Process issues and research needs. *Early Education and Development*, 20(3), 377-401.
- *Shernoff, E. S., & Kratochwill, T. R. (2007). Transporting an evidence-based classroom management program for preschoolers with disruptive behavior problems to a school: An analysis of implementation, outcomes, and contextual variables. *School Psychology Quarterly*, 22(3), 449-472.
- *Shidler, L. (2009). The impact of time spent coaching for teacher efficacy on student achievement. *Early Childhood Education Journal*, 36(5), 453-460.
- Shure, M. B. (1990). *I Can Problem Solve: An interpersonal cognitive problem solving program (preschool)*. Champaign, IL: Research Press.

- *Sibley, A., & Sewell, K. (2011a). Can multidimensional professional development improve language and literacy instruction for young children? *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 263-274.
- Sibley, A., & Sewell, K. (2011b). Elements of multidimensional professional development that make a difference. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(4), 339-344.
- Simon, S., & Sachse, S. (2013). Anregung der Sprachentwicklung durch ein Interaktionstraining für Erzieherinnen. *Diskurs Kindheits- und Jugendforschung*, 4(8), 379-397.
- *Sipp, R. B. (2010). *The efficacy of Emotional and Instructional Support (EIS) training and consultation on Head Start teacher-child interactions*. Unpublished dissertation, University of Nevada, Las Vegas.
- Slavin, R. E. (2008). Perspectives on evidence-based research in education. What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.
- Smith-Jentsch, K. A., Salas, E., & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate. *Journal of Applied Psychology*, 86(2), 377-401.
- *Snyder, J., Low, S., Schultz, T., Barner, S., Mareno, D., Garst, M., et al. (2011). The impact of brief teacher training on classroom management and child behavior in at-risk preschool settings: Mediators and treatment utility. *Journal of Applied Developmental Psychology*, 32(6), 336-345.
- Snyder, P., Hemmeter, M. L., Meeker, K. A., Kinder, K., Pasia, C., & McLaughlin, T. (2012). Characterizing key features of the early childhood professional development literature. *Infants & Young Children*, 25(3), 188-212.
- Sonkoff, J., & Phillips, D. A. (2000). *From neurons to neighborhood*. Washington, D.C.: National Academy Of Sciences.
- Sperry, S. K. (1999). *The effects of three professional development methods on preschool teachers' use of classroom management skills and the social behavior of at-risk preschool children*. Unpublished dissertation, University of Oregon, Eugene.
- *Sprague, K., Glantz, F., Kokopelli, F. G., Raya-Carlton, P., & Schilder, D. (2009). *What is gained from a randomized controlled trial and what is lost? Planning, policy, and fear: Effects and lessons learned from a cluster-randomized trial assessing a community-advocacy program's professional development initiative for early childhood educators*. Paper presented at the Society for Research on Educational Effectiveness Conference, Washington, D.C.
- Staehle, W. H., Conrad, P., & Sydow, J. (1999). *Management – Eine verhaltenswissenschaftliche Perspektive*. München: Vahlen.
- Sterling-Turner, H. E., Watson, T. S., & Moore, J. W. (2002). The effects of direct training and treatment integrity on treatment outcomes in school consultation. *School Psychology Quarterly*, 17(1), 47-77.
- Sterne, J. A. C., Egger, M., & Moher, D. (2008). Addressing reporting bias. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 297-333). Chichester, UK: Wiley-Blackwell.
- Stevens, C. K., & Gist, M. E. (1997). Effects of self-efficacy and goal orientation training on negotiation skill maintenance: What are the mechanisms? *Personnel Psychology*, 50(4), 955-978.
- *Stoiber, K. C., & Gettinger, M. (2011). Functional assessment and positive support strategies for promoting resilience: Effects on teachers and high-risk children. *Psychology in the Schools*, 48(7), 686-706.
- Strickland, D. S. & Riley-Ayers, (2006). Early literacy. Policy and practice in the preschool years. Preschool Policy Brief 10. Retrieved October 25th, 2014 from [//nieer.org/resources/policybriefs/10.pdf](http://nieer.org/resources/policybriefs/10.pdf)
- Strickland, D. S., & Schickedanz, J. (2004). *Learning about print in preschool: Working with letters, words, and beginning links with phonemic awareness*. Newark, DE: International Reading Association.

- Sutton, A. J. (2004). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 435-452). New York, NY: Russell Sage.
- Sylva, K. (2010). Quality in early childhood settings. In K. Sylva, E. Melhuish, P. Sammons, I. Siraj-Blatchford & B. Taggart (Eds.), *Early childhood matters: Evidence from the Effective Pre-school and Primary Education Project* (pp. 70-91). Oxford, UK: Routledge.
- Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology*, 76(6), 759-769.
- Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology*, 90(4), 692-709.
- Textor, M. R. (2012). Erziehungs- und Bildungspläne. Retrieved October 18th, 2013, from <http://www.kindergartenpaedagogik.de/1951.html>
- *Thanasetkorn, P. (2009). The impact of 101s: Guide to positive discipline teacher training on teacher interaction practices, teacher-child relationship quality, school adjustment, and academic outcomes in kindergarten classrooms in Bangkok, Thailand. Unpublished dissertation. Old Dominion University.
- Tietze, W., Lee, H.-J., Bensel, J., Haug-Schabel, G., Aselmeier, M., & Egert, F. (2013). Pädagogische Qualität in Kindertageseinrichtungen und Tagespflegestellen. In W. Tietze, F. Becker-Stoll, J. Bensel, A. G. Eckhardt, G. Haug-Schnabel, B. Kalicki, H. Keller & B. Leyendecker (Eds.), *NUBBEK. Nationale Untersuchung zu Bildung, Betreuung und Erziehung in der frühen Kindheit* (pp. 69-87). Weimar, Berlin: Verlag das netz.
- Torgesen, J. K., & Mathes, P. G. (1998). What every teacher should know about phonological awareness. Retrieved December 10th, 2012, from <http://www.fldoe.org/ese/pdf/phon9872.pdf>
- Tout, K., Isner, T., & Zaslow, M. (2011). Coaching for quality improvement: Lessons learned from Quality Rating and Improvement Systems. Retrieved December 10th, 2012, from <http://www.childtrends.org/wp-content/uploads/2013/05/2011-35CoachingQualityImprovement1.pdf>
- Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). Issues for the next decade of Quality Rating and Improvement Systems. Retrieved May, 8th, 2013, from http://www.childtrends.org/Files//Child_Trends-2009_5_19_RB_QualityRating.pdf
- Tracey, J. B., Tannenbaum, S. I., & Kavanagh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. *Journal of Applied Psychology*, 80(2), 239-252.
- Tremblay, R. (2000). The development of aggressive behavior during childhood: What have we learned in the past century? *International Journal of Behavioral Development*, 24(2), 129-141.
- Tremblay, R. E., Nagin, D. S., Seguin, J. R., Zoccolillo, M., Zelazo, P., Boivin, M., et al. (2004). Physical aggression during early childhood: Trajectories and predictors. *Pediatrics*, 114(1), 43-50.
- Trivette, C. M., Raab, M., & Dunst, C. J. (2012a). Steps to successful professional development in Head Start. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 15(1), 127-134.
- Trivette, C. M., Raab, M., & Dunst, C. J. (2012b). An evidence-based approach to professional development in Head Start classrooms. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 15(1), 41-58.
- *Tyler, K. K. (2009). *The effects of data-driven instruction and literacy coaching on kindergartners' literacy development*. Unpublished dissertation, Indiana University of Pennsylvania, Indiana.
- Tziner, A., Haccoun, R. R., & Kadish, A. (1991). Personal and situational characteristics influencing the effectiveness of transfer of training improvement strategies. *Journal of Occupational Psychology*, 64(2), 167-177.
- U.S. Department of Education. (2006). Improving teacher quality state grants, ESEA Title II, Part A. Non-regulatory guidance. Retrieved December 15th, 2012, from <https://www2.ed.gov/programs/teacherqual/guidance.pdf>

- U.S. Department of Education. (2002). Making a difference in the live of infants and toddlers and their families: The impact of Early Head Start. Volume 1. Final technical report. Retrieved October 15th, 2012, from <http://www.mathematica-mpr.com/~media/publications/PDFs/ehsfinalvol1.pdf>
- *Uttley, C. M., & Horm, D. M. (2008). Mentoring in early childhood professional development: Evaluation of the Rhode Island Child Development Specialist Apprenticeship Program. *Journal of Early Childhood Teacher Education*, 29(3), 237-252.
- Vandell, D. L., & Wolfe, B. (2000). Child care quality: Does it matter and does it need to be improved? Retrieved December 3rd, 2012, from <https://aspe.hhs.gov/ccquality00/ccqual.htm>.
- Varol, F., Farran, D. C., Bilbrey, C., & Vorhaus, E. A. (2012b). Improving mathematics instruction for early childhood teachers: Professional development components that works. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field*, 14(1), 24-40.
- Varol, F., Farran, D. C., Bilbrey, C., Vorhaus, E. A., & Guess Hofer, K. (2012a). Professional development for preschool teachers: Evidence and practice. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field* 15(1), 122-126.
- *Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M., & Amendum, S. (2010). *The targeted reading intervention: A classroom teacher professional development program to promote effective teaching for struggling readers in kindergarten and first grade*. Paper presented at the Society for Research on Educational Effectiveness Conference, Washington, D.C.
- Vestal, A., & Jones, N. A. (2004). Peace building and conflict resolution in preschool children. *Journal of Research in Childhood Education*, 19(2), 131-142.
- von Hippel, A., & Grimm, R. (2010). Qualitätsentwicklungskonzepte in der Weiterbildung Frühpädagogischer Fachkräfte. Weiterbildungsinitiative Frühpädagogische Fachkräfte. WiFF Expertisen, Band 3. Retrieved June, 4th, 2013, from <http://www.weiterbildungsinitiative.de/uploads/media/HippelGrimm.pdf>
- Walker, E., Hernandez, A. V., & Kattan, M. W. (2008). Meta-analysis: Its strenghts and limitations. *Cleveland Clinic Journal of Medicine*, 75(6), 431-439.
- Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational & Organizational Psychology*, 72(3), 351-375.
- Warr, P. B., & Allan, C. (1998). Learning strategies and occupational training. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 83-121). London, UK: Wiley.
- *Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology*, 98(1), 63-74.
- *Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology*, 103(2), 455-469.
- Webster-Stratton, C. (1997). Early intervention for families of preschool children with conduct problems. In M. J. Guralnick (Ed.), *The effectiveness of early intervention* (pp. 429-453). Baltimore, MD: Brookes.
- Webster-Stratton, C. (2000). *The Incredible Years Training Series. Juvenile Justice Bulletin*. Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention.
- Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Weinheim und Basel: Beltz Verlag.
- Weinert, S., Doil, H., & Frevert, S. (2008). Kompetenzmessungen im Vorschulalter. Eine Analyse vorliegender Verfahren. In H.-G. Roßbach & S. Weinert (Eds.), *Kindliche Kompetenzen im Elementarbereich: Förderbarkeit, Bedeutung, Messung* (pp. 89-209). Berlin: Bundesministerium für Bildung und Forschung.

- Weinstock, P., Bos, J., Tseng, F., Rosenthal, E., Ortiz, L., Dowsett, C., et al. (2012). *Evaluation of Program for Infant/Toddler Care (PITC): An on-site training of caregivers. Final Report*. Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Science, U.S. Department of Education.
- Wexley, K. N., & Baldwin, T. T. (1986). Post-training strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal*, 29(3), 503–520.
- Wexley, K. N., & Nemeroff, W. F. (1975). Effectiveness of positive reinforcement and goal setting as methods of management development. *Journal of Applied Psychology*, 60(4), 446–450.
- What Works Clearinghouse. (2011). Procedures and standards handbook. (Version 2.1) Retrieved August 8th, 2013, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf
- What Works Clearinghouse. (2013). Procedures and standards handbook. (Version 3.0) Retrieved November 11th, 2013, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf
- Whitebook, M. (2002). Bachelor's degrees are best: Higher qualifications for pre-kindergarten teachers lead to better learning environments for children. Retrieved March 15th, 2011, from http://www.leg.state.vt.us/PreKEducationStudyCommittee/Documents/NCSL_bachelor_degrees_and_better_learning_pdf_Document.PDF
- Whitebook, M., Gomby, D., Bellm, D., Sakai, L., & Kipnis, F. (2009a). Preparing teachers of young children: The current state of knowledge, and a blueprint for the future. Part I: Teacher preparation and professional development in grades K-12 and in early care and education: Differences and similarities, and implications for research. Retrieved September 24th, 2013, from http://www.irle.berkeley.edu/cscce/wp-content/uploads/2009/01/teacher_prep_1.pdf
- Whitebook, M., Gomby, D., Bellm, D., Sakai, L., & Kipnis, F. (2009b). Preparing teachers of young children: The current state of knowledge, and a blueprint for the future. Part II: Effective teacher preparation in early care and education: Toward a comprehensive research agenda. Retrieved September 24th, 2013, from http://www.irle.berkeley.edu/cscce/wp-content/uploads/2009/01/teacher_prep_2.pdf
- Whitebook, M., & Ryan. (2011). Degrees in context: Asking the right questions about preparing skilled and effective teachers of young children. Preschool Policy Brief 11. Retrieved October 24th, 2012 from <http://nieer.org/resources/policybriefs/23.pdf>
- Whitehurst, G. J. & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, 69(3), 848-872.
- Wilcox-Herzog, A., McLaren, M., Ward, S., & Wong, E. (2013). Results from the Quality Early Childhood Training Program. *Journal of Early Childhood Teacher Education*, 34(4), 335-349.
- Wilson, D. B., & Lipsey, M. W. (2000). Practical meta-analysis. Thousand Oaks, CA: Sage Publications.
- Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, 6(4), 413-429.
- Wirtz, M., & Casper, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Wood, M. M. (1997). *Social competence for young children: An outreach project for inservice training*. Athens, GA: Georgia University.
- Wood, J. & McLemore, B. (2001). Critical components in early literacy – Knowledge of the letter of the alphabet and phonics instruction. *The Florida Reading Quarterly*, 38(2), 1-8.

- Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033). Retrieved June 5th, 2013, from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033.pdf
- Zaslow, M. (2009). Strengthening the conceptualization of early childhood professional development initiatives and evaluations. *Early Education and Development*, 20(1), 1-10.
- Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010a). Towards the identification of features of effective professional development for early childhood educators. Literature review. Retrieved June 5th, 2013, from <http://www2.ed.gov/rschstat/eval/professional-development/literature-review.pdf>
- Zaslow, M., Tout, K., Halle, T., Whittaker, J. V., & Lavelle, B. (2010b). Emerging research on early childhood professional development. In S. B. Neuman & M. L. Kamil (Eds.), *Preparing teachers for the early childhood classroom* (pp. 19-47). Baltimore, MD: Brookes.
- Zellman, G. L., & Perlman, M. (2008). *Child Care Quality Rating and Improvement Systems in five pioneer states: Implementation issues and lessons learned*. Santa Monica, CA: RAND Education & RAND Corporation.
- Zellman, G. L., Perlman, M., Le, V.-N., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar early learning quality rating and improvement system as a tool for improving child-care quality*. Santa Monica, CA: RAND Education & RAND Corporation.
- Zirpoli, T. J., & Melloy, K. J. (1993). *Behavior management: Applications for teachers and parents*. New York, NY: Maximilian.

Appendix A

Coding Scheme (short form)

1. Date [date]: _____
 2. Coder ID [coder_id]: _____
 3. Study ID number [studyid]: _____
 4. Finding Number [findnr]: _____
 5. Author [auth]: _____
 6. Title of Publication [title]: _____
-
7. Publication year [pubyr]: _____
 8. Type of publication [publish]
 - (1) ☐ published (e.g., article, book, chapter)
 - (2) ☐ unpublished (e.g., report, dissertation)
 9. Type of study [st_type]: _____ page(s): _____
 - (1) ☐ program with control/comparison group
 - (2) ☐ program without control/comparison group
 10. Study design [st_design]: _____ page(s): _____
 - (1) ☐ experimental
 - (2) ☐ quasi-experimental
 - (3) ☐ correlational
 - (4) ☐ one group longitudinal
 - (5) ☐ single subject design
 11. Mechanism of allocation [assign]: _____ page(s): _____
 - (1) ☐ random assignment (at individual level)
 - (2) ☐ random assignment (at group/center level)
 - (3) ☐ non-random assignment, but control group matched to treatment group
 - (4) ☐ non-random, no matching prior to treatment
 - (5) ☐ assignment to condition not described or vague
 - (99) ☐ study without experimental design / no assignment to conditions
 12. Target group of intervention [target]: _____ page(s): _____
 - (1) ☐ preschool/ kindergarten teachers
 - (2) ☐ K-12 teachers
 - (3) ☐ other (e.g., caregivers, nurses)
 - (99) ☐ no teachers or caregivers / cannot determine
 13. Number of participants(teachers at beginning): _____ page(s): _____

experimental group 1: _____ [exgr1]	control group 1: _____ [congr1]
experimental group 2: _____ [exgr2]	experimental group 3: _____ [exgr3]
total sample: _____ [tosamp]	
 14. Kind of program [program]: _____ page(s): _____
 - (1) ☐ in-service teacher training
 - (2) ☐ pre-service teacher training
 - (3) ☐ other: _____ [program3]
 - (99) ☐ cannot determine / not reported

15. Kind of training design: {code 1 or 9} page(s): _____
☐ workshop [workshop]
☐ course (e.g., University course) [course]
☐ online training/ web-mediated professional development [online]
☐ community of learners/ community of practice [community]
☐ Program with onsite support [onsite]

16. Kind of training/onsite support [meth_onsite] : page(s): _____
 (1) ☐ coaching
 (2) ☐ consulting
 (3) ☐ other: _____ [meth_onsite3]
 (98) ☐ no onsite support method
 (99) ☐ cannot determine / not reported

17. Teacher performance measurement [teachper_measure] page(s): _____
 (1) ☐ external measurements (e.g., observation, rating)
 (2) ☐ self reporting/ self evaluation (e.g., teacher questionnaire)
 (98) ☐ no performance measurement
 (99) ☐ cannot determine / not reported

18. Child outcome measures [childoutcome]: page(s): _____
 (1) ☐ cognitive development or measures of school readiness (M, SD, N or ES, N)
 (2) ☐ social development (M, SD, N or ES, N)
 (3) ☐ cognitive, school readiness and social development (M, SD, N or ES, N)
 (98) ☐ no child outcomes
 (99) ☐ cannot determine / not reported

19. Quality rating/Quality measurement [quality_measure] page(s): _____
☐ quality measurement {code 1 or 9}
 Instrument: _____ [quality_instr]

20. Provision of appropriate outcome measurement data: {code 1 or 9}
☐ external performance measurements (M, SD, N or ES, N) [performance] page(s): _____
☐ quality ratings (M, SD, N or ES, N) [quality] page(s): _____
☐ child outcomes (M, SD, N or ES, N) [childout] page(s): _____

21. Relevance for analysis [relevance]: (1) ☐ relevant (2) ☐ not relevant
 Notes:

☐ contact the author

Appendix B

Coding Scheme (Extended Form) CHILD OUTCOMES

1. Date [date]: _____
2. Coder ID [coder_id]: _____
3. Study ID [studyid]: _____
4. Finding number [findnr]: _____
5. Treatment number [treatnr]: _____
6. Author [auth]: _____
7. Publication year [pubyr]: _____
8. Publication country [pubcountry]: _____
9. Publication type [pubtype]:
(1) ☐ published (e.g., article, book) (2) ☐ unpublished (e.g., report, dissertation)

CODING AT STUDY LEVEL

CONTEXT

10. Context [context] page(s): _____
- Please, select the context in which the study is conducted.
- (1) ☐ center-based care
- (2) ☐ early childhood education (preschool, kindergarten or pre-k programs)
- (3) ☐ early intervention program for disadvantaged (e.g., Head Start, CARE, Abecedarian)
- (4) ☐ elementary school
- (5) ☐ family day care
- (6) ☐ other or mixture: _____ [context_o]
- (99) ☐ cannot determine

TEACHER SAMPLE

11. Number of participants (teacher at beginning): page(s): _____
- experimental group 1: _____ [exgr1t] control group 1: _____ [congr1t]
- experimental group 2: _____ [exgr2t] experimental group 3: _____ [exgr3t]
- total sample: _____ [tosamp]
12. Sample of caregivers / teachers consist of [cons_teach] page(s): _____
- Please, tick all that fits. {code 1=yes or 0=no}
- ☐ Lead Teacher/Head Teachers
- ☐ Assistant Teachers
- ☐ Center Directors
- ☐ Specialists (e.g., speech pathologists/ literacy coaches)
- ☐ other: _____ [cons_teach_o]

13. Qualification/certification of caregivers / teachers in sample [exp_teach] page(s): _____
Please, tick all that fits. {code 1=yes or 0=no}

- ☐ none
- ☐ high school
- ☐ vocational or technical school course or college course / GED (general education diploma)
- ☐ associate's degree
- ☐ bachelor's degree
- ☐ master's degree
- ☐ Ph.D.
- ☐ other: _____
- ☐ cannot determine / not reported

14. Dropout rate of teacher [drop_teach] page(s): _____
Please, describe the percentage of dropout from pretest to posttest
_____ % of dropout
Calculation: _____

Please code the following dichotomous items for the teacher sample homogeneity
{1=yes/0=no /9 missing}

15. All teacher/caregiver in the sample have a university degree (BA/MA/Ph.D.)? page(s): _____
☐ yes ☐ no
16. All teacher/caregiver in the sample are lead or assistant teachers? page(s): _____
☐ yes ☐ no
17. The overall dropout rate is below 25 %? page(s): _____
☐ yes ☐ no

CHILD SAMPLE

18. Number of participants (children at beginning): {9999 missing} page(s): _____

experimental group 1: _____ [exgr1c] control group 1: _____ [congr1c]
experimental group 2: _____ [exgr2c] experimental group 3: _____ [exgr3c]
total sample: _____ [tosampc]

19. Age group of children in the overall sample [age_child]
page(s): _____

- (1) ☐ children 0- 3 years of age
- (2) ☐ children 3 - 6 years of age
- (3) ☐ children 0 - 6 years of age
- (4) ☐ other _____
- (99) ☐ cannot determine

20. Dropout rate of children [drop_chil] page(s): _____
Please, describe the percentage of dropout from pretest to posttest
_____ % of dropout
Calculation: _____

Please code the following dichotomous items for the child sample homogeneity
{1=yes/0=no /9 missing}

21. Are children at risk / disadvantaged children the majority of the sample? page(s): ____
☐ yes ☐ no

22. Are children with special needs/handicaps included in the sample? page(s): ____
☐ yes ☐ no

23. Are children below 3 years part of the sample? page(s): ____
☐ yes ☐ no

24. Does the sample consist of children that are only in homogeneous age groups?
(e.g., only preschool children / only kindergarten children) page(s): ____
☐ yes ☐ no

25. Percentage of families with low SES in the child sample [lowses_child] page(s): ____
{code 1=yes or 9=no}
☐ more than 50 % of the child sample has a low SES

FOCUS OF THE PROFESSIONAL DEVELOPMENT PROGRAM

26. Name of the Training /Professional Development Program [programe]:

27. Focus of Training [kindprogram]: page(s): ____
☐ formal education (part of a college/university training to receive a degree e.g., AA, BA, MA)
☐ semi-formal education (e.g., CDA, ECE course)
☐ informal education (e.g., workshop, community college course, online training)

28. Target group of training intervention [targetint]:
page(s): ____
(1) ☐ preschool/ kindergarten teachers
(2) ☐ K-12 teachers
(3) ☐ nurses/ (unspecified) caregivers/educators
(4) ☐ others _____
(100) ☐ no teachers or caregivers / cannot determine

29. Curriculum implementation as major focus of intervention? {code 1=yes or 0=no}
☐ yes ☐ no

CODING AT TREATMENT LEVEL

REPEAT / COPY THE CODING AT TREATMENT LEVEL FOR EACH TREATMENT

30. Treatment ID

Study ID	Finding NR	Treatment NR

[treatid]:

SAMPLE : TEACHER / CAREGIVER

31. Gender of caregivers / teachers [gender_teach] / [gender_teach_per] page(s): _____

Treatment	Control
Percentage (females):	Percentage (females):

32. Age of caregivers / teachers [age_teach] page(s): _____

Treatment	Control
Mean:	Mean:
SD:	SD:
Minimum:	Minimum:
Maximum:	Maximum:

33. Years of experience of caregivers / teachers [exp_teach] page(s): _____

Treatment	Control
Mean:	Mean:
SD:	SD:
Minimum:	Minimum:
Maximum:	Maximum:

34. Percentage of degree of caregivers / teachers²¹ [exp_teachp] page(s): _____

Percentage:	Treatment	Control
University degree (BA, MA, PhD)		
Some training (AA,CDA)		
No Training (high school or less)		

²¹ University Degree (Ph.D./MA/BA) // Some Training (AA/CDA) // no training (teacher assistant)

SAMPLE : CHILDREN

35. Gender of children [gender_child]

page(s): _____

Treatment	Control
Percentage (females):	Percentage (females):

36. Age of children [age_child]

page(s): _____

Treatment	Control
Mean:	Mean:
SD:	SD:
Minimum:	Minimum
Maximum:	Maximum:

37. Socio-economical status of children

page(s): _____

Percentage:	Total	Treatment	Control
High SES			
Middle SES			
Low SES			

38. Children with risk factors in sample

page(s): _____

- (1) ☐ children at risk due to SES
- (2) ☐ children at risk due to developmentally delays
- (3) ☐ children at risk due to immigration background/ethnicity
- (4) ☐ children at risk due to low birth weight
- (5) ☐ children at risk due to multiple factors
- (6) ☐ no children at risk in the sample

39. Children with disadvantages/at risk in sample (percentage)

Percentage:	Treatment	Control
Children at risk due to SES		
Children at risk due to developmentally delays		
Children at risk due to immigration background		

Professional Development/Training

CONTENT OF THE PROFESSIONAL DEVELOPMENT (WHAT)

40. Topic of the professional development program / training [toptraining] page(s): ____
Please, tick the one that fits best!

- (1) ☐ First aid, health & hygiene training
- (2) ☐ Math & science
- (3) ☐ Language and literacy
- (4) ☐ Social, emotional & behavioral aspects/ Sociability / Peer-Interaction
- (5) ☐ Nutrition & sport
- (6) ☐ Quality Improvement (e.g., structure or process quality)
- (7) ☐ Curriculum or model (for implementation)
- (8) ☐ other: _____
- (99) ☐ cannot determine

41. Contents of the professional development program / training page(s): ____

Please, tick all that fit! {code 1=yes or 0=no}

- ☐ curriculum, model or program (e.g., Recognition & Response)
 - ☐ child development (e.g., language development)
 - ☐ child assessment / child observation
 - ☐ working with parents/parent involvement
 - ☐ working with children at risk / second language learners (general focus)
 - ☐ early intervention & scaffolding (e.g., concrete didactical/fostering strategies for children at risk)
 - ☐ classroom environment & material provision (e.g., books selection)
 - ☐ classroom management & improvement of daily routines (e.g., mixture of group interactions)
 - ☐ teaching styles (e.g., child initiated or developmental appropriateness)
→ pedagogical focus
 - ☐ didactical principles (e.g., providing & planning of didactical lessons for children)
→ educational delivery focus and principles of providing new contents to children
 - ☐ improvement of teacher behavior / interaction with children (e.g., teacher language)
→ improvement of interactions between teacher and child
 - ☐ Others:
-
-

42. Expected outcome of training {code 1=yes or 0=no}

Please, tick all that fits!

Teacher outcomes [focustea]:

page(s): _____

- ☐ teacher knowledge
- ☐ teacher attitudes / orientations / awareness
- ☐ teacher performance (skills, interactions, behavior)

Quality outcomes [focusqual]:

page(s): _____

- ☐ general child care quality (e.g., ECERS, ITERS, CLASS)
- ☐ specific child care quality of one educational domain (e.g., ELLCO)
- ☐ target child quality (e.g., EAS)

Program quality outcomes [focusfid]:

page(s): _____

- ☐ fidelity check / program integrity of the professional development program (for teacher)
- ☐ fidelity / program integrity of the educational curriculum implemented (for children)

Child outcomes [focuschi]:

page(s): _____

- ☐ cognitive development & school readiness
- ☐ language & literacy
- ☐ math & science
- ☐ social, emotional and behavioural
- ☐ other: _____ [focuschio]

DELIVERY OF THE PROFESSIONAL DEVELOPMENT (HOW)

43. Kind of training design: {code 1=yes or no=9}

page(s): _____

Please, tick all that fit!

- ☐ workshop /institute (e.g., 1 day seminar with no follow-up lessons) [workshop]
- ☐ course (e.g., University course / long-term training with more lessons) [course]
- ☐ online training/ web-mediated professional development (includes web coaching) [online]
- ☐ community of learners/ community of practice [community]
- ☐ Program with onsite support (e.g., coaching, mentoring, supervision) [onsite]

44. Duration of PD /Training [pd_duration]

page(s): _____

Please, describe the time frame in which the whole training should be delivered

(e.g., June 2009 – May 2010 = 12 months)

_____ total duration in months (end – beginning)

Calculation: _____

45. Kind of onsite support [meth_onsite] : {code 1 or 9} page(s): _____
 Please, tick all that fit!
☐ coaching [coaching]
☐ consulting [consulting]
☐ mentoring [mentoring]
☐ supervision [supervision]
☐ video feedback [video]
☐ individual lessons (no specified method reported) [individual]
☐ other: _____ [onsite_o]

DOSIS PD-SESSION (workshop/ institute/ course / online/ onsite)

46. Length of each PD-session (workshop/institute/course) [pd_length_session]²² page(s): _____
 Please, describe
 - the length of each session in hours (e.g., 30 min = 0,5 / 20min = 0,33 / 15 min=0,25)
 - the frequency / number of sessions (place (use the mean if available / e.g., 6 - 8=7)
 - the rhythm of sessions provided (use codes below)
 1= weekly / 2=biweekly / 3=monthly / 4=quarterly / 5=sporadic /
 99=☐ cannot determine

PD-Type	Length in hours [x_length_session]	Number of Sessions [x_num_session]	Rhythm of sessions [x_rhy_session]
Workshop / initiative [w]			
Course [c]			
Online lessons [o]			
Community of learners [cl]			
Onsite support [on]			

47. Total length of PD /Training (workshop/institute/course) [pd_length] page(s): _____

Please, calculate the total length PD-Sessions
 (e.g., course with 6 lessons á 4 h + 1 introduction workshop (8 hours) = 32 h)

_____ total length of PD-sessions (in hours)
 Calculation: _____

²² Calculate 8 h for a full day workshop/ 4 h for a half day workshop / 30 h for a college or university course

Please code the following dichotomous items for training characteristics
{1=yes/0=no /9 missing}

48. PD/Training consists only of onsite support? page(s): ____
☐ yes ☐ no

49. Mix delivery methods are used in the PD/Training (e.g., onsite & course)? page(s): ____
☐ yes ☐ no

50. The total duration of the PD/Training is > 1 year? page(s): ____
☐ yes ☐ no

51. The total length of the PD/Training is more than 30 hours? page(s): ____
☐ yes ☐ no

52. PD is a scale-based training /uses scale-based feedback? page(s): ____
☐ yes ☐ no

53. In the PD the whole team participate (collective participation)? page(s): ____
☐ yes ☐ no

TRAINERS OF THE PROFESSIONAL DEVELOPMENT (WHO)

54. Number of Professional Development Provider in the study page(s): ____
 ____ Number of PD-Providers/Trainers

55. Person delivery the professional development/training page(s): ____
 (1) ☐ researcher/faculty members/faculty colleagues
 (2) ☐ graduate students
 (3) ☐ professor (e.g., college courses/ university course)²³
 (4) ☐ professional coaches/trainers/
 (5) ☐ multipliers
 (6) ☐ specialists (e.g., speech & language therapists)
 (99) ☐ cannot determine / not reported

56. Education/ Highest degree of professional development providers/trainers page(s): ____
 (1) ☐ none
 (2) ☐ some training / qualification (high school/ college course/ associate degree)
 (3) ☐ university degree (BA / MA / Ph.D.)
 (99) ☐ cannot determine / not reported

57. Experience of professional development providers {1=yes/0=no /9 missing} page(s): ____

Providers are described as experienced in coaching/mentoring/providing PD?

☐ yes ☐ no

58. Child care experience of PD providers {1=yes/0=no /9 missing} page(s): ____

Providers has teaching experience in early childhood classrooms?

☐ yes ☐ no

²³ If the PD is a college course, please mark professor.

Mark also „3 university degree“ at item 57, „yes, experienced“ at item 58, „no multiplier“ at item 60, and „no special training“ at 61.

59. Multiplier approach {1=yes/0=no /9 missing} page(s): _____

Trainers are multipliers (receiving the training once and should spread their knowledge afterwards)?

☐ yes ☐ no

60. Training of Providers/ Trainers {1=yes/0=no /9 missing} page(s): _____

61.

Trainers/providers receive special training before they start to teach/coach?

☐ yes ☐ no

62. Comment on training of trainers

How does the training of professional development providers/trainers/coaches looks like?

Comment:

TRAINING TRANSFER FACTORS

63. Foreknowledge/pre-course of participants {1=yes/0=no /9 missing} page(s): _____

The participants had to do a pre-course/pre-training before the PD)?

☐ yes ☐ no

64. Group of participants {1=yes/0=no /9 missing} page(s): _____

The participants had the (free) choice to participate in this course/training/PD?

☐ yes ☐ no

65. Material working environment {1=yes/0=no /9 missing} page(s): _____

Participants receive material to implement the new curriculum/strategies?

☐ yes ☐ no

EDUCATIONAL CURRICULUM FOR CHILDREN

***ONLY IF THE PD FOCUS ON THE IMPLEMENTATION OF A MODEL/CURRICULUM!!!

66. Length of each educational curriculum [pd_length_session]²⁴ page(s): _____

Please, describe

- the length of each session in hours (e.g., 30 min = 0,5 / 20min = 0,33 / 15 min=0,25)
- the frequency / number of sessions (place (use the mean if available / e.g., 6 - 8=7)
- the rhythm of sessions provided (use codes below)
1=daily / 2= weekly / 3=biweekly / 4=monthly / 5=quarterly / 6=sporadic / 99=missing

Length in hours [x_length_session]	Number of Sessions [x_num_session]	Rhythm of sessions [x_rhy_session]

67. Dosis of educational curriculum [pd_length] page(s): _____

Please, calculate the dosis of the educational curriculum

_____ total dosis of educational curriculum (in hours)

Calculation: _____

68. Duration of educational curriculum [curriculum_duration] page(s): _____

Please, describe the time frame in which the whole training should be delivered
(e.g., June 2009 – May 2010 = 12 months)

_____ total duration in months (end – beginning) /Calculation²⁵:

69. Focus of educational curriculum {1=yes/0=no /9 missing} page(s): _____

Does the educational curriculum focuses on the fostering of children with disadvantages/being at risk?

☐ yes ☐ no

²⁴ Calculate 8 hours for a full day workshop 4 h for a half day workshop/30 h for a college or university course

²⁵ For one school year calculate 9 months

METHODOLOGY

GENERAL METHODOLOGICAL CHARACTERISTICS TEACHER

70. Is there are a comparison/control group? [comp_group] page(s): _____

- (1) ☐ program with control/comparison group(s)
 (2) ☐ program without control/comparison group

71. Treatment of control group [treat_group] page(s): _____

Is there a real control group?

- (1) ☐ yes, control group doesn't receive a treatment / business as usual (e.g., wait list)
 (2) ☐ yes, control group receive a placebo treatment that doesn't correlate with the exp. Treatment (placebo e.g., control group=drawing with watercolours / treat group= language program)
 (3) ☐ no, only different treatment levels are compared in this study
 (4) ☐ no, no control group available (e.g., one group longitudinal design)

(5)

72. Amount of treatment & comparison groups used in study [n_treat_group] {99=missing}

Number of treatment groups: (e.g., treatment groups with different training levels)	N=
Number of control groups (receiving nothing): (e.g., waitlist or placebo)	N=

73. Assignment to treatment/control group teacher [assign_group] page(s): _____

- (1) ☐ random assignment (at individual level)
 (2) ☐ random assignment (at group level)
 (3) ☐ random assignment (at center/site level)
 (4) ☐ matching at individual level
 (5) ☐ matching at group/center level
 (6) ☐ non-random assignment, groups of convenience, condition not described or vague

GENERAL METHODOLOGICAL CHARACTERISTICS CHILDREN

74. Assignment to treatment/control group children [assign_group_child] page(s): _____

- (1) ☐ random assignment (at individual level)
 (2) ☐ random assignment (at group level)
 (3) ☐ random assignment (at center/site level)
 (4) ☐ matching at individual level
 (5) ☐ matching at group/center level
 (6) ☐ non-random assignment, groups of convenience, condition not described or vague

75. Evaluation outcome domains {code 1=yes or 0=no}

child outcomes [child_out]:

page(s): _____

- ☐ language
- ☐ literacy
- ☐ maths
- ☐ science
- ☐ social-emotional/behavioral development
- ☐ cognition

76. Program fidelity

page(s): _____

	professional development	child care curriculum
Fidelity in %		
Fidelity rated as... 1=high 0=not with high fidelity		

OVERALL COMMENT:

[illegible]

STATISTICAL DATA / EFFECT SIZE CALCULATION

For data & effect size extraction, please use the excel-sheet.