Erfassung und Veränderung der allgemeinen Unterrichtsqualität im Rahmen der Lehrerfortbildungsstudie „Lernen mit Plan“

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Dipl. Päd. Sarah Werth

aus Friedberg (Hessen)

Tübingen

2014

Tag der mündlichen Prüfung:	27.11.2014
Dekan:	Professor Dr. rer. soc. Josef Schmid
1. Gutachter:	Professor Dr. phil. Ulrich Trautwein
2. Gutachter:	Professor Dr. phil. Benjamin Fauth

DANKSAGUNG

Die vorliegende Arbeit entstand im Rahmen meines Promotionsstipendiums an der Universität Tübingen in der Abteilung empirische Bildungsforschung und pädagogische Psychologie (EBPP) und wurde ermöglicht durch ein Promotionsstipendium des Landes Baden-Württemberg und die Förderung des zugrundliegenden Forschungsprojekts durch das Bundesministerium für Bildung und Forschung (BMBF).

Mein großer Dank gilt meinen Gutachtern Prof. Dr. Ulrich Trautwein und Prof. Dr. Benjamin Fauth. Vor allem möchte ich an dieser Stelle aber Dr. Wolfgang Wagner danken, der die Entstehung der Arbeit intensiv betreute und meinen niemals endenden Fragen mit einer ebenfalls nie endenden Geduld begegnete und tapfer versuchte, mir die Grundlagen des statistischen Denkens nahezubringen. Auch allen anderen Mitarbeiterinnen und Mitarbeitern der EBPP möchte ich an dieser Stelle herzlich danken für fast vier Jahre des gemeinsamen Austauschs und zahlreicher schöner Momente. Ganz besonderer Dank gilt an dieser Stelle auch den hochgeschätzten und liebgewonnenen Kooperationspartnern aus der Technischen Universität Darmstadt, ohne deren gute Zusammenarbeit sich die Lehrerfortbildung „Lernen mit Plan“ nie hätte realisieren lassen und mit denen zahlreiche arbeitsame Abende und Wochenenden doch auch immer einige heitere Momente bereithielten.

Mit Abgabe dieser Dissertationsschrift blicke ich auf eine Zeit zurück, in der ich intensiv miterlebt und festgestellt habe, dass die Realisierung anspruchsvoller Forschungsarbeiten den Universitätsmitarbeitern viel Idealismus, Enthusiasmus und Engagement abverlangt und dass alle, die sich unter den teilweise schwierigen Arbeitsbedingungen behaupten, meine höchste Bewunderung haben. Ich habe darüber hinaus aber auch gemerkt, dass mein Herz mehr für die Gestaltung schulischer Praxis als für die Erforschung selbiger schlägt. Meine im Rahmen des wissenschaftlichen Werdegangs erworbenen Kenntnisse werde ich daher nach Vollendung dieser Dissertation wie bereits im vergangenen Jahr in der Zusammenarbeit mit den Schulen des Kreises Viersen einsetzen. Ich freue mich daher darauf, in Zukunft stärker mit der Rezeption wissenschaftlicher Befunde befasst zu sein als mit dem Verfassen derselben. Dennoch möchte ich mich an dieser Stelle auch bei Prof. Dr. Hermann Josef Abs bedanken, der meinen wissenschaftlichen Werdegang immer gefördert hat und bis zuletzt versuchte, mich für die Weiterarbeit in der Wissenschaft zu motivieren.

Abschließend möchte ich nun noch den Personen danken, die mich auf meinem bisherigen Weg begleitet und mir mit Rat und Tat zur Seite gestanden haben: meine Familie, mein Partner und meine guten Freundinnen und Freunde, die für meine Sorgen, Nöte und Höhenflüge immer ein offenes und—zum Glück—auch kritisches Ohr hatten.

ZUSAMMENFASSUNG

Die Qualität von Unterricht hat einen bedeutenden Einfluss auf das Lernen und die Motivation von Schülerinnen und Schülern. In zahlreichen Forschungsarbeiten wurden daher bereits Merkmale von erfolgreichem Unterricht identifiziert, diese zu Dimensionen zusammengefasst und Maßnahmen konzipiziert, um die Unterrichtsqualität zu verbessern.

Gleichzeitig lässt sich jedoch feststellen, dass es auch im Zusammenhang mit der Erfassung von Unterrichtsqualität noch einige ungeklärte Fragen gibt. So ist bisher nicht untersucht, ob die in Querschnittserhebungen vielfach monierte geringe Übereinstimmung von Lehrer- und Schülerurteilen auch in Längsschnitterhebungen bestätigt werden kann. Hierbei ist insbesondere von Interesse, ob die geringe Übereinstimmung zwischen Lehrer- und Schülerurteilen durch situative Einflüsse zum jeweiligen Erhebungszeitpunkt erklärt werden kann und sich die Übereinstimmung zwischen Lehrer- und Schülereinschätzungen der Unterrichtsqualität durch die Aggregation mehrerer Messzeitpunkte verbessert. Darüber hinaus stellt sich die Frage, ob Fortbildungsmaßnahmen, die in das Unterrichtsgeschehen eingreifen, nicht nur Auswirkungen auf trainierte Aspekte des Unterrichts haben, sondern möglicherweise auch „Nebenwirkungen“ auf andere, nicht adressierte Aspekte der Unterrichtsgestaltung, die sich dann wiederum auf die Leistungsentwicklung der Schülerinnen und Schüler auswirken.

Die vorliegende Dissertation entstand im Kontext des Forschungsprojekts „Lernen mit Plan“, das zum Ziel hatte, mithilfe einer Lehrerfortbildung im Rahmen des Mathematikunterrichts die Selbstregulationskompetenz von Schülerinnen und Schülern der fünften Jahrgangsstufen an Haupt- und Werkrealschulen in Baden-Württemberg zu fördern.

Anhand von drei Teilstudien, die auf den Daten der Studie „Lernen mit Plan“ basieren, untersucht die vorliegende Dissertation daher drei Forschungsfragen:

Teilstudie 1 untersucht die Forschungsfrage, inwieweit Lehrer- und Schülerratings der Unterrichtsqualität zu den drei Erhebungszeitpunkten der Studie messzeitpunktspezifische und messzeitpunktübergreifende Varianzanteile aufweisen. Außerdem wird in dieser Teilstudie untersucht, ob die Übereinstimmung zwischen Lehrer- und Schülerurteilen der Unterrichtsqualität durch Kontrolle der messzeitpunktspezifischen Varianz höher ist als die bisher nur in Querschnittstudien ermittelte Lehrer-Schülerübereinstimmung.

Teilstudie 2 geht der Fragestellung nach, inwieweit sich aus Lehrer- und Schülersicht die Teilnahme der Lehrkräfte an der Fortbildung „Lernen mit Plan“ nicht nur auf die Umsetzung der fokalen Trainingsinhalte (selbstregulationsspezifische Unterrichtsqualität), sondern auch

auf die nicht trainierten Aspekte der Unterrichtsqualität (allgemeine Unterrichtsqualität) ausgewirkt hat.

In Teilstudie 3 wird aufbauend auf den Befunden von Teilstudie 2 geprüft, ob die Fortbildung wie erwartet eine Verbesserung der Mathematikleistung der Schülerinnen und Schüler bewirkt hat und ob dieser Effekt durch die Umsetzung der fokalen Trainingsinhalte oder durch die mögliche fortbildungsbedingte Veränderung der allgemeinen Unterrichtsqualität vermittelt wird.

Die Ergebnisse der Studien weisen zum einen darauf hin, dass die bisher in der Forschungsliteratur benannten Diskrepanzen zwischen Lehrer- und Schülerurteilen bei der Erfassung von Unterrichtsqualität auch auf messzeitpunktspezifische Einflüsse zurückzuführen sind. Darüber hinaus lassen die Befunde der vorliegenden Dissertation die Annahme zu, dass eine Lehrerfortbildung zur Förderung der Selbstregulation von Schülerinnen und Schülern auch Veränderungen auf nicht trainierter Aspekte der Unterrichtsqualität haben kann, diese nicht intendierten Veränderungen jedoch keinen Einfluss auf die als Outcome untersuchte Mathematikleistung der Schülerinnen und Schüler haben.

Die vorliegende Dissertation stellt damit einen wichtigen Beitrag zur weiteren Erforschung der Eigenschaften von Lehrer- und Schülerurteilen zur Erfassung der Unterrichtsqualität dar. Darüber hinaus leistet die Dissertation einen wichtigen Beitrag zur Aufklärung der Wirkmechanismen von Lehrerfortbildungen als Maßnahme zur Förderung der Selbstregulation von Schülerinnen und Schülern und der Auswirkungen solcher Fortbildungen auf die Mathematikleistung dar.

In der vorliegenden Dissertation werden zunächst die in den Teilstudien erarbeiteten Forschungsfragen theoretisch hergeleitet. Daran anschließend werden die drei Teilstudien vorgestellt. Die Ergebnisse der Teilstudien werden abschließend in einer Gesamtdiskussion zusammengeführt und erörtert. Im letzten Kaptiel werden auch die Grenzen der vorliegenden Dissertation und die Implikationen für die jeweiligen Forschungsfelder und die Praxis dargestellt.

INHALT

1

Einleitung und theoretischer Rahmen der Arbeit

1 Einleitung und theoretischer Rahmen der Arbeit

Lehrerfortbildungen haben das Potenzial zur Veränderung der Unterrichtspraxis und können Schülerleistungen positiv beeinflussen (Borko, 2004; Yoon, Duncan, Lee, Scarloss & Shapley, 2007). Damit sind sie eine wichtige Maßnahme zur Professionalisierung von Lehrkräften (Seidel & Shavelson, 2007) und zur unterrichtsbezogenen Qualitätsentwicklung an Schulen (Borko & Putnam, 1995; Darling-Hammond, Wei, Andree, Richardson & Orphanos, 2009). Trotz dieser zentralen Bedeutung von Lehrerfortbildungen ist die Befundlage zu ihrer Wirksamkeit in vielen Bereichen derzeit als eher „jung“ einzustufen und zu vielen Forschungsfragen fehlen noch gesicherte Erkenntnisse.

So wird bei Lehrerfortbildungen, sofern sie aufgrund diverser forschungsmethodischer Schwierigkeiten überhaupt realisiert werden können, häufig nur überprüft, ob die Teilnahme an der Lehrerfortbildung einen Einfluss auf die Leistungsentwicklung der Schülerinnen und Schüler zeigt. Untersuchungen, ob die Fortbildungsteilnahme tatsächlich zu einer Veränderung der Unterrichtspraxis führt, werden hingegen nur selten durchgeführt und den ermittelten Ergebnissen fehlt es aufgrund der zumeist kleinen Stichproben (Yoon et al., 2007) zudem an Generalisierbarkeit. Darüber hinaus wurde bisher nicht untersucht, wie sich Lehrerfortbildungen auf nicht adressierte Aspekte der allgemeinen Unterrichtsqualität auswirken. In der vorliegenden Studie wurden daher im Rahmen der Lehrerfortbildung „Lernen mit Plan“, welche die Verbesserung der Selbstregulationsförderung im Mathematikunterricht zum Ziel hatte, drei verschiedene Aspekte untersucht. Zum einen wurde in Studie 1 zunächst der Forschungsfrage nachgegangen, ob die bereits vielfach durch die Forschung belegten Differenzen zwischen Lehrer- und Schülerurteilen zur Unterrichtsqualität in Längsschnittstudien durch die Aggregation verschiedener Messzeitpunkte und die Separation von konstanten und variablen Wahrnehmungsanteilen verringert werden können. In Studie 2 wurde untersucht, inwieweit die Lehrerfortbildung „Lernen mit Plan“ sich nicht nur auf die trainierten Inhalte auswirkt, sondern auch auf untrainierte Aspekte der allgemeinen Unterrichtsqualität. In Studie 3 wurde abschließend überprüft, inwieweit die Lehrerfortbildung Effekte auf die Mathematikleistung der Schülerinnen und Schüler zeigt und ob diese Effekte durch die Umsetzung der Fortbildungsinhalte mediiert wrrden oder ob diese durch Veränderungen der allgemeinen Unterrichtsqualität vermittelt werden.

Im theoretischen Rahmen der Dissertation wird in den Abschnitten *1.1 Erfassung von Unterrichtsqualität aus Lehrer- und Schülersicht* und *1.2 Veränderung der allgemeinen Unterrichtsqualität im Rahmen einer Lehrerfortbildung zur Förderung der Selbstregulation*

von Schülerinnen und Schülern im Mathematikunterricht der theoretische Hintergrund der Dissertation dargestellt. Aus diesen theoretischen Grundlagen werden unter 1.3 die der Arbeit zugrundeliegenden Fragestellungen abgeleitet. Im Anschluss daran werden in den Kapiteln 2 bis 4 die im Rahmen dieser Dissertation angefertigten Studien dargestellt, in denen die zuvor hergeleiteten Fragestellungen dezidiert erarbeitet werden.

Abschließend werden in Kapitel 5 die Ergebnisse der drei Studien zusammengefasst und ein Ausblick auf noch offene Forschungsfragen gegeben.

1.1 Erfassung von Unterrichtsqualität aus Lehrer- und Schülersicht

Unterrichtsqualität umfasst alle Interaktionen zwischen Lehrkräften und Schülerinnen bzw. Schülern, die die Leistungs- und Motivationsentwicklung der Schülerinnen und Schüler beeinflussen (Caroll, 1963; Weinert, Schrader & Helmke, 1989), und wird als bedeutsam für deren Lernen und Mitarbeit angesehen (Hattie, 2008; Scheerens & Bosker, 1997; Seidel & Shavelson, 2007). Zahlreiche Forschungsarbeiten haben sich daher einerseits der Untersuchung der Dimensionalität von Unterrichtsqualität gewidmet und Modelle zur Beschreibung der Unterrichtsqualität entwickelt. Im Abschnitt 1.1.1 sollen daher zunächst das dieser Dissertation zugrundeliegende theoretische Modell der Unterrichtsqualität und die darin postulierten Zusammenhänge zwischen den Dimensionen der Unterrichtsqualität und der Leistungs- und Motivationsentwicklung dargestellt werden.

Eine weiterer Forschungsstrang im Bereich der Unterrichtsforschung hat sich intensiv mit der Erfassung von Unterrichtsqualität auseinander gesetzt. Ausgehend von den Befunden dieses Forschungsstrangs, der sich überwiegend auf Studien mit querschnittlichem Design bezieht, untersucht die vorliegende Dissertation mögliche Auswirkungen von längsschnittlichen Erhebungen auf die Erfassung der Unterrichtsqualität. Im Abschnitt 1.1.2 soll daher zunächst auf die Vor- und Nachteile von Lehrer- und Schülerfragebögen als Instrumente zur Erfassung von Unterrichtsqualiät eingegangen werden. Daran anschließend wird im Abschnitt 1.1.3 auf Befunde zur Übereinstimmung von Lehrer- und Schülerurteilen in querschnittlichen Designs eingegangen. Abschließend werden vor diesem Hintergrund im Abschnitt 1.1.4 mögliche Implikationen aktueller Befunde aus Studien mit Beobachterratings für die Lehrer-Schüler-Übereinstimmung bei längsschnittlicher Erfassung der Unterrichtsqualität erörtert.

1.1.1 Unterrichtsqualität als zentrale Grundlage für schulisches Lernen[1]

Die Definition von Unterrichtsqualität in der vorliegenden Arbeit entspricht dem Modell der drei Basisdimensionen der Unterrichtsqualität nach Klieme, Schümer und Knoll (2001). Der Ansatz dieses Modells ist dabei, die Vielzahl an Unterrichtsmerkmalen zu integrieren, die im Rahmen der Unterrichtsforschung charakterisiert wurden und deren Wirksamkeit in zahlreichen Studien nachgewiesen wurde (vgl. Hattie, 2008; Seidel & Shavelson, 2007).

Dieses Modell (Abb. 1) kategorisiert die Unterrichtsmerkmale in drei Basisdimensionen der Unterrichtsqualität, nämlich Klassenführung, kognitive Aktivierung und Schülerorientierung. Die Basisdimension Klassenführung, unter der sich zahlreiche Aspekte der direkten Instruktion (Rosenshine, 1970) wiederfinden, beschreibt inwieweit es einer Lehrperson gelingt, einen störungsfreien und gut organisierten Unterrichtsverlauf zu gewährleisten. Wichtige Kriterien sind dabei sowohl, dass es der Lehrkraft gelingt, den notwendigen Überblick über Klassenzimmerprozesse zu behalten, als auch die zügige und nach Möglichkeit präventive Abhandlung von Disziplinproblemen (Klieme, Lipowsky, Rakoczy & Ratzka, 2006; Ophardt & Thiel, 2008; Waldis, Grob, Pauli & Reusser, 2010). Kognitive Aktivierung, welche die zweite Dimension des Modells darstellt, beschreibt die Kompetenz der Lehrkraft, Schülerinnen und Schüler im Unterricht zu einer bewussten und aktiven Auseinandersetzung mit den Unterrichtsinhalten anzuregen (Klieme et al., 2006). Mögliche Aktivitäten der Lehrkraft in dieser Dimension umfassen einerseits eine das Vorwissen der Schülerinnen und Schüler aktivierende Aufgabenstellung und andererseits die Ermunterung der Schülerinnen und Schüler, eigene Lösungswege zu suchen, diese zu begründen und selbstständig zu überprüfen. Das Finden des optimalen Anspruchsniveaus ist dabei von zentraler Bedeutung, da die Schülerinnen und Schüler zwar gefordert, jedoch nicht überfordert werden sollten (Klieme et al., 2006; Kunter & Voss, 2011). Die Schülerorientierung als dritte Dimension der Unterrichtsqualität beschreibt, wie sehr sich der Unterricht auch an den Bedürfnissen der Schülerinnen und Schüler orientiert und damit motivierend auf diese wirkt (Klieme et al., 2006). Auch hier ist es wichtig, dass bei den Schülerinnen und Schülern kein Gefühl der Überforderung entsteht—allerdings wird in dieser Dimension nicht die Passung der Aufgabenschwierigkeit berücksichtigt, sondern vielmehr inwieweit die Lehrkraft die Lernenden im Lernprozess anleitet und begleitet. Dabei ist sowohl ein respektvoller und geduldiger Umgang mit den Lernenden wichtig als auch das Erkennen

[1] Dieser Abschnitt stellt eine leicht modifizierte Form des entsprechenden Abschnitts aus Studie II dar. Die Formulierung wurde hierbei nicht abgeändert, da es aus Sicht der Autorin keine bessere Möglichkeit gibt, diese theoretischen Erkenntnisse stringent darzustellen und zusammenzuführen.

und Beheben von individuellen Verständnisschwierigkeiten (Klieme et al., 2006; Kunter, 2005). Die Leistungsentwicklung wird dabei gemäß Klieme et al. (2006; 2009) insbesondere durch die erste und zweite Basisdimension, Klassenführung und Kognitive Aktivierung, beeinflusst (Klieme et al., 2006; Klieme et al., 2009). Die dritte Dimension, Schülerorientierung, hat hingegen zunächst einen Einfluss auf die Motivationsentwicklung und über diese einen indirekten Effekt auf die Leistungsentwicklung (Klieme et al., 2006; Klieme et al., 2009; Lipowsky et al., 2009). Im weiteren Verlauf der vorliegenden Arbeit werden die drei Basisdimensionen unter dem Begriff *allgemeine Unterrichtsqualität* zusammengefasst.

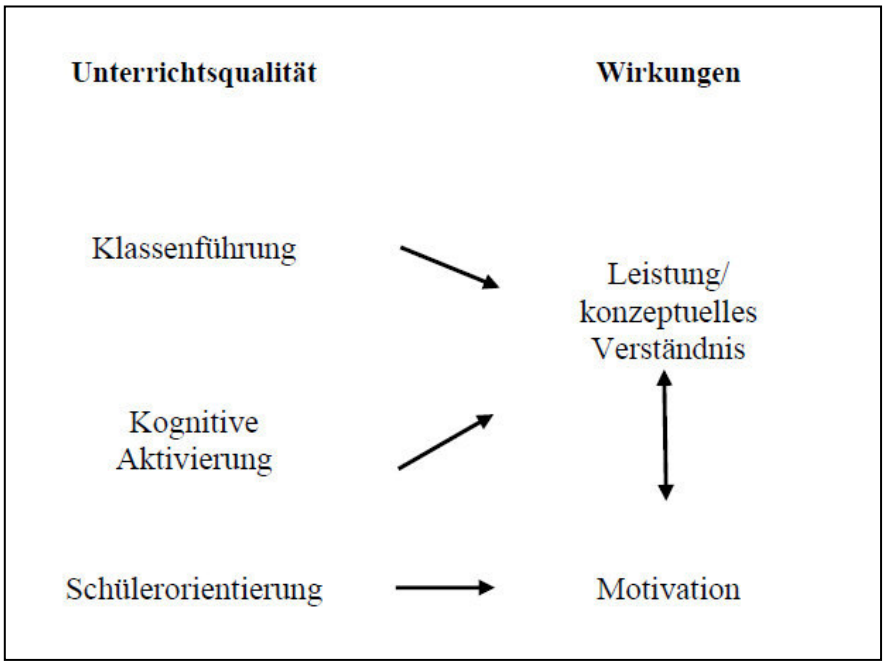

Abb. 1: Modifiziertes Modell der Basisdimensionen guten Unterrichts und deren vermuteter Wirkung nach Klieme et al. (2006)

1.1.2 Erfassung von Unterrichtsqualität durch Lehrer- und Schülerfragebögen

Die allgemeine Unterrichtsqualität ist von essentieller Bedeutung für die kognitive Entwicklung und die Motivation der Schülerinnen und Schüler (Klieme et al., 2009; Seidel & Shavelson, 2007) und daher Ziel zahlreicher großangelegter Studien (z.B. Kane & Staiger, 2012). Zur Erfassung der Unterrichtsqualität haben sich derzeit drei Datenquellen etabliert, nämlich Lehrer- und Schülerfragebogen und Unterrichtsbeobachtung durch geschulte Beobachter (Clausen, 2002). Jede dieser Quellen hat dabei aus forschungsmethodischer Sicht sowohl Vor- als auch Nachteile (Clausen, 2002). Zumeist wird Unterrichtsqualität jedoch durch fragebogenbasierte Lehrer- und Schülerurteile erfasst, da diese kostengünstiger sind als Beobachterratings (Desimone, Smith & Frisvold, 2009; Lanahan, McGrath, McLaughlin, Burian-Fitzgerald & Salganik, 2005). Als Vorteil von Lehrerurteilen werden die theoretischen und meist auch praktischen Kenntnisse verschiedener Unterrichtsmethoden seitens der Lehrkräfte angeführt. Auch können Lehrkräfte in der Beurteilung von Unterricht—abhängig von der Anzahl ihrer Dienstjahre—auf langjährige Erfahrung im Strukturieren und Durchführen von Unterrichtsstunden zurückgreifen (Porter, 2002). Darüber hinaus kann als Vorteil von Lehrerurteilen zur Erfassung von Unterrichtsqualität auch die unter anderen von Clausen (2002) ermittelte hohe Übereinstimmung zwischen Lehrkräften und Beobachtern bei der Beurteilung des eigenen Unterrichts angeführt werden. Außerdem erwiesen sich Lehrerurteile in der Studie von Clausen (2002) auch als prädiktiv für das Lernen von Schülerinnen und Schülern. Als Einschränkung der Validität von Lehrerurteilen wird u.a. von Desimone et al. (2009) angeführt, dass die Genauigkeit der Ratings abhängig von individuellen Hintergrundmerkmalen der Lehrkräfte wie Unterrichtserfahrung und Ausbildungsgrad ist (Kunter et al., 2005). Auch weisen Forschungsbefunde von bspw. Porter (2002) darauf hin, dass die Lehrkräfte bei der Beurteilung des eigenen Unterrichts möglicherweise von eigennützigen Motiven oder auch eigenen Idealvorstellungen von Unterrichtsgestaltung beeinflusst sein könnten (Wubbels, Brekelmans & Hooymayers, 1992). Die These der eigennützigen, d.h. eher beschönigenden Unterrichtsbeurteilung durch Lehrkräfte, konnte jedoch von anderen Wissenschaftlern wie z.B. Kunter und Baumert (2006) nicht bestätigt werden.

Als Vorteil von Schülerurteilen zur Erfassung von Unterricht wird angeführt, dass Schülerinnen und Schüler als “Experten” für die Beurteilung von Unterricht gelten können, da sie täglich mit dem Unterricht verschiedener Lehrkräfte konfrontiert sind und daher zur

Urteilsbildung auf „Vergleichswerte“ zurückgreifen können (Clausen, 2002; De Jong & Westerhof, 2001). Ein weiterer Vorteil von Schülerurteilen ist deren im Vergleich zu Lehrer- und Beobachterratings höhere Prädiktionskraft für die Leistungs- und Motivationsentwicklung (Kunter & Baumert, 2006; Kunter et al., 2005). Nicht zuletzt aufgrund dieser Eigenschaft hat sich der Einsatz von auf Klassenebene aggregierten Schülerurteilen zur Erfassung der Unterrichtsqualität und zur Vorhersage der Motivations- und Leistungsentwicklung im Rahmen der empirischen Bildungsforschung durchgesetzt (De Jong & Westerhof, 2001). Jedoch weisen wissenschaftliche Untersuchungen auch auf Nachteile des Einsatzes von Schülerurteilen zur Erfassung der Unterrichtsqualität hin. So deuten Forschungsbefunde darauf hin, dass die Wahrnehmung von Schülerinnen und Schülern durchaus deutlich von anderen Datenquellen wie Lehrerurteilen und Beobachterurteilen abweicht und damit eine eigenständige Wahrnehmungsperspektive darstellt (Clausen, 2002; Kunter & Baumert, 2006). So wirken sich auch im Fall von Schülerurteilen auf Individualebene Hintergrundmerkmale der Schülerinnen und Schüler wie beispielsweise Geschlecht, Alter und Schulleistung auf die Beurteilung von Unterricht aus (Aleamoni, 1999; Gentry, Gable & K., 2002; Kunter & Baumert, 2006). Auf Klassenebene haben darüber hinaus die Strenge der Lehrkraft bei der Notenvergabe und die Beliebtheit der Lehrkraft einen Einfluss auf die Beurteilung des Unterrichts (Aleamoni, 1999; Greenwald, 1997). Ein weiterer Nachteil von Schülerurteilen zur Erfassung der Unterrichtsqualität ist die in der Forschungsliteratur erwähnte hohe Interkorrelation von Schülerurteilen zu verschiedenen Konstrukten, die dadurch begründet wird, dass Schülerurteile durch den globalen Eindruck, den die Schülerinnen und Schüler von der Lehrkraft haben, beeinflusst werden (Kunter & Baumert, 2006; Wagner, 2008).

In Folge muss davon ausgegangen werden, dass sowohl Lehrer- als auch Schülerurteile einen perspektivenspezifischen Varianzanteil aufweisen, der berücksichtigt werden sollte, wenn diese Quellen zur Beurteilung von Unterricht und anschließend zur Prädiktion von Motivations- und Leistungsentwicklung herangezogen werden (Clausen, 2002; Desimone et al., 2009; Kunter & Baumert, 2006).

1.1.3 Übereinstimmung von Lehrer- und Schülerurteilen zur Erfassung der Unterrichtsqualität in Querschnittserhebungen

Hinsichtlich der Übereinstimmung zwischen Lehrer- und Schülerurteilen von Unterricht haben Forschungsarbeiten eine schwache bis moderate Korrelationen festgestellt. Dabei ist die Höhe der Übereinstimmung zwischen den beiden Perspektiven sowohl von den

Merkmalen der Klasse als auch von den jeweilig zu beurteilenden Aspekten der Unterrichtsqualität abhängig (Clausen, 2002; Desimone et al., 2009; Kunter & Baumert, 2006). Die beispielweise von Clausen (2002) berichteten Korrelationen lagen dabei zwischen $-0.28 < r < 0.42$ und Kunter und Baumert (2006) fanden Faktorkorrelationen zwischen $0.09 < r < 0.64$. Auch hinsichtlich der konstruktspezifischen Unterschiede des Übereinstimmungsgrades zwischen Lehrer- und Schülerurteilen erweisen sich die Forschungsbefunde als konsistent: so wurde sowohl bei Kunter und Baumert (2006) als auch bei Clausen (2002) eine niedrigere Übereinstimmung von $0.09 < r < 0.24$ für hochinferente Konstrukte wie *Aufgabenauswahl* und *Autonomieunterstützung* gefunden. Die höchste Übereinstimmung wurde hingegen für niedriginferente Konstrukte wie *ineffiziente Klassenführung* (r =0.64; Kunter & Baumert, 2006) und *repetitives Üben* (r = .42; Clausen, 2002) gefunden. Darüber hinaus berichtete Clausen (2002), dass die Korrelation zwischen verschiedenen Konstrukten innerhalb der jeweiligen Perspektive höher ist als die Übereinstimmung zwischen gleichen Konstrukten, die aus verschiedenen Perspektiven beurteilt wurden. Allerdings erwiesen sich dabei die Korrelationen zwischen den verschiedenen Konstrukten für die Schülerperspektive als deutlich höher als für die Lehrerperspektive.[2]

Vor dem Hintergrund dieser Forschungsbefunde schlussfolgerten und belegten bspw. Desimone et al. (2009), dass sowohl Lehrer- als auch Schülerurteile von individuellen Hintergrundmerkmalen beeinflusst und damit verzerrt werden und eine statistische Kontrolle der betreffenden Hintergrundmerkmale zu einer höheren Übereinstimmung zwischen den Perspektiven führt. Gemäß dieser Schlussfolgerung könnten beide Perspektiven ohne Kontrolle dieser Hintergrundmerkmale nicht als valide Indikatoren für Unterrichtsqualität angesehen und eingesetzt werden. Im Gegensatz dazu postulieren bspw. Clausen (2002) und Kunter und Baumert (2006), dass die Abweichungen zwischen den beiden Perspektiven nicht nur auf Messfehler zurückgeführt werden können, sondern dass es sich hierbei um eine perspektivenspezifische Validität handelt. Diese Annahme führen sie auf eigene Forschungsbefunde zurück, die zeigen, dass sich auch die Faktorstruktur und die konvergente und diskriminante Vailidität zwischen Lehrer- und Schülerurteilen unterscheiden. In Folge interpretieren Lehrkräfte und Schülerinnen bzw. Schüler Items mit demselben Wortlaut möglicherweise unterschiedlich und kommen dementsprechend zu unterschiedlichen Einschätzungen bzgl. der Unterrichtsqualität.

[2] Dieser Absatz findet sich in modifizierter Form auch in Studie I wieder.

Abschließend kann festgestellt werden, dass—resultierend aus den Befunden zur Spezifität der Wahrnehmungsperspektiven—zwar weiterhin untersucht wird, welche Instrumente sich zur Erfassung welcher Aspekte der Unterrichtsqualität am besten eignen (z.B. Kane & Staiger, 2012). Dennoch werden diese Erkenntnisse bei der Konzeption von Studien aufgrund der organisatorischen Rahmenbedingungen in der empirischen Bildungsforschung nur unzureichend berücksichtigt. So wird aus Kostengründen und bedingt durch die eingeschränkte Größe der Stichproben noch immer häufig ausschließlich auf Schülerfragebögen zurückgegriffen ohne diese Einschränkung in der Interpretierbarkeit der Befunde zu erörtern. Des Weiteren wird in großangelegten Studien, die bspw. den Effekt einer Bildungsreform auf die Unterrichtsqualität ermitteln, häufig eine wiederholte Messung der Unterichtsqualität vorgenommen. Im folgenden Abschnitt soll daher auf die Lehrer-Schüler-Übereinstimmung im Längsschnitt eingegangen werden.

1.1.4 Übereinstimmung von Lehrer- und Schülerurteilen zur Erfassung der Unterrichtsqualität in Längsschnitterhebungen[3]

Obwohl sich zusammenfassend sagen lässt, dass die Eigenschaften von Lehrer- und Schülerurteilen zur Unterrichtsqualität schon intensiv untersucht wurden, so ist doch festzustellen, dass sich diese Befunde ausschließlich auf Querschnittserhebungen beziehen und es nach Erkenntnis der Autorin derzeit keine Befunde zur Übereinstimmung zwischen Lehrer- und Schülerurteilen in Längsschnitterhebungen gibt. Zwar werden in längsschnittlichen Studien häufig Lehrer- und Schülerfragebögen eingesetzt um Schlussfolgerungen z.B. über die Effekte von Unterrichtsqualität auf die Leistungsentwicklung oder aber die Effekte von Interventionen auf die Unterrichtsqualität zu ermitteln. Dennoch ist nur wenig darüber bekannt, wie sich die Übereinstimmung von Lehrer- und Schülerurteilen im Rahmen von mehrmaligen Messungen verhält.

Aktuelle Forschungsergebnisse zur Reliabilität von längsschnittlichen Beobachterratings aus dem MET-Projekt (Kane & Staiger, 2012) deuten jedoch darauf hin, dass Unterrichtsratings sowohl messzeitpunktüberdauernde Varianzanteile als auch messzeitpunktspezifische Varianzanteile beinhalten. So aggregierten Kane und Staiger (2012) zur Ermittlung der Größe von verschiedenen Varianzkomponenten die Beobachterratings von vier Messzeitpunkten im MET-Projekt und zerlegten diese dann in rater- und messzeitpunktspezifische Varianzkomponenten.

---

[3] Abschnitt 1.1.4 ist eine wörtliche Übersetzung eines Abschnitts aus Studie I. Die Formulierung wurde hierbei nicht abgeändert, da es aus Sicht der Autorin keine bessere Möglichkeit gibt, diese theoretischen Erkenntnisse stringent darzustellen und zusammenzuführen.

Da Lehrer- und Schülerurteile möglicherweise stärker durch Messwiederholung beeinflusst sein könnten als geschulte Beobachter, stellt sich jedoch die Frage, ob diese Ergebnisse auch auf Lehrer- und Schülerratings übertragbar sind. So könnten die Ratings von Lehrern- und Schülerinnen bzw. Schülern, die zuvor im Gegensatz zu Beobachtern keine Raterschulung erhalten haben, über die Zeit vertrauter mit den Instrumenten werden (Koziol & Burns, 1986; Lievens, Reeve & Heggestad, 2007). Die zunehmende Vertrautheit mit den Fragen könnte über die Zeit zu einer reliableren Urteilsbildung durch Lehrkräfte und Schülerinnen bzw. Schüler führen. Andererseits wäre es möglich, dass die wiederholte Erhebung und die damit verbundene Beanspruchung der Lehrkräfte und Schülerinnen bzw. Schüler zu einer Verringerung der Testmotivation führt (Lanahan et al., 2005).

Befunde von Koziol und Burns (1986) deuten darauf hin, dass die Genauigkeit von Lehrerratings sich verbessert, wenn Lehrkräfte eine Beurteilung des eigenen Unterrichts mehrmals vornehmen. So fanden sie heraus, dass die Korrelation zwischen Lehrer- und Beobachterratings bei wiederholter Messung höher war als zum ersten Messzeitpunkt. Auch Clausen (2002) äußert die Annahme, dass Messwiederholung dabei helfen könnte, den Messfehler durch situationsspezifische Einflüsse in Unterrichtsratings zu reduzieren. Unter Bezugnahme auf Weinstein (1985) nimmt er hierbei an, dass Schülerinnen und Schüler bei der Einschätzung von Unterrichtsqualität mitunter stark durch besonders einprägsame Situationen in den vergangenen Tagen oder Wochen beeinflusst sein könnten. In Folge wäre es möglich, dass Schülerratings der Unterrichtsqualität bei mehrmaliger Erfassung durchaus variieren könnten.

Eine Erfassung der Unterrichtsqualität zu mehr als zwei Messzeitpunkten könnte dementsprechend dazu führen, eine genauere Annäherung an die durchschnittliche und von situativen Einflüssen bereinigte Einschätzung der Unterrichtsqualität zu erhalten und damit möglicherweise eine höhere Übereinstimmung zwischen Lehrer- und Schülerratings zu erzielen. Diese Annahme findet außerdem Unterstützung durch die Befunde von Marsh und Grayson (1994), die in längsschnittlichen, konfirmatorischen Faktorenanalysen messzeitpunktspezifische Varianzkomponenten gezielt untersuchen und zeigen können, dass sich dort durch die Aggregation von mehreren Messzeitpunkten situations- bzw. messzeitpunktspezifische Varianzkomponenten vom Messfehler separieren lassen.

Davon ausgehend, dass auch Lehrer- und Schülerratings sich in eine variable und eine konsistente Komponente zerlegen lassen, wäre anzunehmen, dass die variable Komponente der Unterrichtswahrnehmung u.a. stark durch perspektivenspezifische Referenzperioden beeinflusst wird (Clausen, 2002; Lanahan et al., 2005; Weinstein, 1985). So könnten z.B.

klassenübergreifende Eindrücke (wie z.B. die Schwierigkeit der letzten Klassenarbeit oder die Schwierigkeit der aktuellen thematischen Einheit) durchaus zu einer messzeitpunktspezifischen, unterschiedlichen Wahrnehmung zwischen Lehrkräften und Schülerinnen bzw. Schülern beitragen. Es wäre daher durchaus möglich, dass eine von der variablen Komponente bereinigte (aggregierte) konsistente Komponente eine höhere Lehrer-Schüler-Übereinstimmung aufweist als die unzerlegten Lehrer- und Schülerurteile zu den einzelnen Messzeitpunkten.

Diese Befunde und Überlegungen legen nahe, dass sich durch die Aggregation mehrerer Messzeitpunkte auch Lehrer- und Schülerratings in sowohl eine variable als auch eine zeitstabile Komponente zerlegen lassen und dass die Übereinstimmung zwischen Lehrer- und Schülerurteilen zwischen der extrahierten zeitstabilen Komponente höher ist als die aus querschnittlichen Erhebungen berichtete Übereinstimmung zwischen den Perspektiven. Eine empirische Überprüfung dieser Überlegungen wurde bisher jedoch nicht durchgeführt.

Im Anschluss an die Darstellung der Definition von Unterrichtsqualität und der aktuellen Befunde zur Erfassung von Unterrichtsqualität durch Lehrer- und Schülerratings soll im folgenden Teil der Arbeit auf die Veränderung von Unterrichtsqualität im Rahmen von Lehrerfortbildungen eingegangen werden.

1.2 Veränderung der allgemeinen Unterrichtsqualität im Rahmen einer Lehrerfortbildung zur Förderung der Selbstregulation von Schülerinnen und Schülern im Mathematikunterricht

In Anbetracht der großen Bedeutung von Unterrichtsqualität für die Leistungs- und Motivationsentwicklung von Schülerinnen und Schülern stellt sich die Frage, inwieweit Unterrichtsqualität durch Lehrerfortbildungen beeinflusst werden kann bzw. im Rahmen von Lehrerfortbildungen beeinflusst wird. In den folgenden Abschnitten soll daher zunächst unter 1.2.1 auf Lehrerfortbildungen als Maßnahme der Professionalisierung von Lehrkräften eingegangen werden. Dabei werden u.a. forschungsmethodologische Schwierigkeiten bei der Erfassung der Effektivität von Lehrerfortbildungen thematisiert. Unter 1.2.2 werden anschließend verschiedene Ansätze zur Gestaltung von Lehrerfortbildungen zur Förderung der Selbstregulation von Schülerinnen und Schülern und Befunde zu deren Wirksamkeit dargestellt. Unter 1.2.3 werden diese Befunde aufgegriffen und intensiv erörtert, in welcher Form möglicherweise auch nicht trainierte Aspekte der Unterrichtsqualtiät durch solche Lehrerfortbildungen beeinflusst werden. Abschließend werden unter 1.2.4 die Effekte von selbstregulationsfördernden Lehrerfortbildungen auf die Mathematikleistung der Schülerinnen

und Schüler als häufig in diesen Lehrerfotbildungen untersuchtes Outcome dargestellt. Der Schwerpunkt dieses letzten Abschnitts liegt auf der Hinterfragung, wie mögliche Effekte solcher Fortbildungen auf die Mathematikleistung vermittelt werden.

1.2.1 Lehrerfortbildungen als Maßnahme zur Professionalisierung von Lehrkräften

Forschungsergebnisse belegen, dass effektive Lehrerfortbildungen das Potenzial zur Veränderung der Unterrichtspraxis beinhalten und eine positive Wirkung auf Schülerleistungen aufweisen (Borko, 2004; Yoon et al., 2007). Lehrerfortbildungen sind damit ein wichtiger Baustein der angestrebten Professionalisierung von Lehrkräften (Seidel, Rimmele & Prenzel, 2005; Seidel & Shavelson, 2007) und gelten als zentrale Maßnahme zur weiteren unterrichtsbezogenen Qualitätsentwicklung an Schulen (Borko & Putnam, 1995; Darling-Hammond et al., 2009).

Unabhängig von der nachgewiesenen, potentiellen Effektivität von Lehrerfortbildungen, erweist sich der Transfer von Fortbildungsinhalten in die Unterrichtspraxis jedoch als ein komplexer und anspruchsvoller Prozess (Lipowsky, 2010). Voraussetzung für einen gelungenen Transfer ist, dass Lehrkräfte die Fortbildungsinhalte annehmen und in die Gestaltung ihres Unterrichts integrieren (Fullan & Stiegelbauer, 1991). Zahlreiche Studien, deuten jedoch darauf hin, dass dieser Transfer häufig ausbleibt bzw. nur teilweise gelingt. Beispielsweise ermittelten Garet et al. (2011) in einer groß angelegten Studie, in deren Rahmen die teilnehmenden Lehrkräfte von externen Trainern und unterstützt durch Coaching-Sitzungen ein bzw. zwei Jahre trainiert wurden, wesentlich geringere Effekte als von den Forschern erwartet. Des Weiteren deuten Befunde von Goldschmidt und Phelps (2007) darauf hin, dass auch die Nachhaltigkeit von Fortbildungserfolgen nicht vorausgesetzt werden kann. In einer umfangreichen Studie zur Veränderung des fachlichen und pädagogischen Wissens durch eine Lehrerfortbildung stellten sie fest, dass zum zweiten Messzeitpunkt ermittelte Lernzuwächse auf Seiten der Teilnehmer bereits zum dritten Messzeitpunkt wieder abgesunken waren.

Aber nicht nur die schwere Veränderbarkeit des Lehrerhandelns stellt die Forschung zu Lehrerfortbildungen vor Herausforderungen. So führen nicht zuletzt organisationale und forschungsmethodische Schwierigkeiten dazu, dass der Großteil der zur Erforschung von Fortbildungsmaßnahmen durchgeführten Studien geringe Stichprobengrößen umfasst und vielfach kein Design realisiert werden kann, welches längsschnittliche Analysen im Vergleich zu einer Kontrollgruppe ermöglicht. So ermittelten Yoon et al. (2007) in einer Metaanalyse

zum Einfluss von Lehrerfortbildungen auf Schülerleistungen unter 1343 Studien lediglich neun, welche über ein quasiexperimentelles Design oder ein randomisiertes Kontrollgruppendesign verfügten. Diese wiederum umfassten Stichprobengrößen von lediglich 5 bis 44 Lehrkräften, was die Generalisierbarkeit der Befunde dieser Studien einschränkt. Effekte der Fortbildung wurden in der Regel auf der Schülerebene ermittelt, ohne die geschachtelte Natur der Daten zu berücksichtigen. In Folge fällt eine durchgeführte Signifikanztestung zu liberal aus. Zudem kann die Einschätzung durch die Lehrkräfte als eine wichtige Evaluationsebene aufgrund der fehlenden Teststärke kaum systematisch berücksichtigt werden.

Ein weiterer Schwachpunkt zahlreicher Studien ist darüber hinaus ihr häufig eingeschränkter Fokus auf nur ein mögliches Outcome der Fortbildung, d.h. entweder die Einstellung der Lehrkräfte zur Fortbildung oder aber ihr Wissenszuwachs, die Umsetzung der Inhalte im Unterricht oder die Veränderung der Schülerleistung. Selten hingegen werden diese möglichen Outcomes gleichzeitig zur Evaluation der Wirksamkeit einer Fortbildung herangezogen. Im Gegensatz dazu empfiehlt das Modell zur Trainingsevaluation von Kirkpatrick (1996), auf der ersten Ebene die Erwartung und Einstellung der Teilnehmer gegenüber der Fortbildung, auf der zweiten Ebene den Wissenszuwachs der Teilnehmer, auf der dritten Ebene die tatsächliche Verhaltensänderung bzw. Umsetzung der Fortbildungsinhalte und schließlich auf der vierten Ebene die Erreichung der Ziele der jeweiligen Fortbildung zu untersuchen. Darüber hinaus geht Kirkpatrick (1996) davon aus, dass sich die Outcomes auf den verschiedenen Ebenen gegenseitig bedingen und sich auf jeder der möglichen Untersuchungsebenen relevante Informationen zur Aufklärung der Wirkmechanismen von Fortbildungen identifizieren lassen. Auch wenn mittlerweile empirisch untermauerte Zweifel an der gegenseitigen Abhängigkeit der vier Ebenen voneinander geäußert wurden (Alliger, Tannenbaum, Bennett & Traver, 1997), stellt dieses Modell einen klaren konzeptuellen Rahmen zur Evaluation der Effekte und der Aufklärung von Wirkmechanismen von Lehrerfortbildungen dar. Zur Ermittlung der Fortbildungseffekte auf einer einzelnen Ebene ist der Einbezug der anderen Ebenen zwar nicht notwendig, kann jedoch bei Ausbleiben eines Fortbildungseffekts von großer Erklärungskraft sein. Vor diesem Hintergrund stellt neben geringen Stichprobengrößen auch die bis auf wenige Ausnahmen (z.B. Krammer, Ratzka, Klieme, Pauli & Reusser, 2006) kaum realisierte systematische Trainingsevaluation auf allen vier Ebenen eine Herausforderung für die Lehrerfortbildungsforschung dar.

Nachdem in diesem Abschnitt Erkenntnisse zu Einflussfaktoren bei der Konzeption und methodologische Herausforderungen bei der Evaluation von Lehrerfortbildungen im

Allgemeinen dargestellt wurden, soll im folgenden Abschnitt auf Lehrerfortbildungen eingegangen werden, welche die Selbstregulationsförderung von Schülerinnen und Schülern im Unterricht zum Ziel haben.

1.2.2 Lehrerfortbildungen zur Förderung der Selbstregulation von Schülerinnen und Schülern

Nicht zuletzt durch den nachweisbar positiven Effekt auf die akademische Leistung und die Motivation (Zimmerman, 2001; Zimmerman & Bandura, 1994), ist das selbstregulierte Lernen zunehmend in den Fokus wissenschaftlichen Interesses gerückt. Selbstreguliertes Lernen wird dabei in Anlehnung an Zimmerman (2008) als zyklischer und proaktiver Prozess verstanden. Dieser Prozess umfasst Elemente wie Zielsetzung, Planung, den selektiven Einsatz von Strategien und das Überwachen des eigenen Lernprozesses durch den Schüler und hat zum Ziel, bestimmte Lernergebnisse zu erreichen.

Eine Vielzahl von Studien hat sich mit der Förderung des selbstregulierten Lernens in verschiedenen Kontexten beschäftigt (Dignath & Büttner, 2008). Dabei konnte gezeigt werden, dass die Effektivität der Vermittlung von Selbstregulation besonders in Kombination mit Fachinhalten zu einer Verbesserung der Selbstregulation, aber auch der fachlichen Leistung führte (Labuhn, Zimmerman & Hasselhorn, 2010; Otto, 2007; Perels, Gürtler & Schmitz, 2005).

Um nun eine möglichst frühzeitige Förderung der Selbstregulation zu gewährleisten, kommt der Frage, auf welche Weise die Selbstregulationsförderung Bestandteil des schulischen Unterrichts werden kann, eine besondere Bedeutung zu. Hierbei rückt automatisch die Lehrkraft und ihr Vermögen in den Blick, den Unterricht effektiv zu gestalten und alltagsrelevante Kompetenzen wie das selbstregulierte Lernen zu vermitteln (Dignath & Büttner, 2008). Forschungsbefunde deuten—wenn auch begleitet von den gleichen forschungsmethodischen Schwierigkeiten wie andere Lehrerfortbildungsevaluationen—darauf hin, dass eine Förderung der Selbstregulation der Schülerinnen und Schüler durch eine Lehrerfortbildung erfolgreich umgesetzt werden kann. So führten bspw. Rozendaal, Minnaert und Boekaerts (2005) eine selbstregulationsbezogene Fortbildung mit Lehrkräften aus 14 Klassen durch, in deren Rahmen sie anhand von Clusteranalysen eine signifikante Verbesserung des selbstregulierten Verhaltens derjenigen Klassen fanden, deren Lehrkräfte angaben, das Programm in überdurchschnittlichem Maße umzusetzen.

Darüber hinaus wurden jedoch auch Studien durchgeführt, die darauf hinweisen, dass die Förderung der Selbstregulation im Unterricht durch Lehrkräfte weniger effektiv ist als die

Selbstregulationsförderung der Schülerinnen und Schüler durch externe Trainer (Dignath & Büttner, 2008; Komorek, Bruder, Collet & Schmitz, 2007; Otto, 2007).

Diese Studien deuten darauf hin, dass—um den Multiplikatoreneffekt der Lehrkräfte zu nutzen—besondere Trainingsansätze für Lehrkräfte und Schülerinnen bzw. Schüler notwendig sind. So wird als eine wirkungsvolle Maßnahme in diesem Zusammenhang genannt, den Lehrkräften die Wichtigkeit der eigenen Umsetzung selbstregulatorischer Prinzipien zu vermitteln, z.B. über die Zielorientierung, die Planung und die Reflexion des Unterrichtsgeschehens (Klug, Ogrin, Keller, Ihringer & Schmitz, 2011). Ein weiterer Anknüpfungspunkt für die Steigerung der Effektivität von Lehrerfortbildungen ist auch der Ansatz, Materialien für ein Trainingsprogramm zu erstellen und sie den Lehrkräften zur Verfügung stellen, damit diese die Materialien in ihrem regulären Unterricht einsetzen. Einen solchen Ansatz verwenden z.B. de Corte, Verschaffel und van De Ven (2001) sowie Mokhlesgerami, Souvignier, Rühl und Gold (2007).

Vor dem Hintergrund dieser Forschungsbefunde lässt sich feststellen, dass hinsichtlich der Vermittlung der Selbstregulation über Lehrkräfte noch Forschungsbedarf besteht. So könnte die bisher ermittelte Überlegenheit von Selbstregulationstrainings durch externe Trainer (Dignath & Büttner, 2008) darauf hindeuten, dass in den bisher untersuchten Lehrerfortbildungen die Fortbildungsinhalte entweder nicht in ausreichender Gründlichkeit vermittelt oder erlernt wurden oder aber der Transfer der Inhalte in die Praxis noch verbessert werden könnte. Der folgende Abschnitt stellt daher Annahmen zu den Wirkmechansimen von Lehrerfortbildungen zur Förderung der Selbstregulation dar.

1.2.3 Effekte von Lehrerfortbildungen zur Förderung der Selbstregulation auf fokale Trainingsinhalte und (nicht-trainierte) Aspekte der allgemeinen Unterrichtsqualität[4]

Zahlreiche Studien, welche die Förderung der Selbstregulation von Schülerinnen und Schülern durch externe Trainer zum Ziel hatten, weisen darauf hin, dass die ökologische Validität dieser Forschungsbefunde eingeschränkt ist, da man schlecht abschätzen könne, ob vergleichbare Effekte eintreten würden, wenn stattdessen Lehrkräfte das Training mit den Schülerinnen und Schülern durchgeführt hätten (u.a. Brunstein & Glaser, 2011; Schünemann, Spörer & Brunstein, 2013). Studien hingegen, welche die Selbstregulation der Schülerinnen und Schüler tatsächlich mithilfe einer Lehrerfortbildung förderten, weisen wiederum auf

[4] Dieser Abschnitt entspricht weitgehend einem Abschnitt aus Studie II. Die Formulierung wurde hierbei nicht abgeändert, da es aus Sicht der Autorin keine bessere Möglichkeit gibt, diese theoretischen Erkenntnisse stringent darzustellen und zusammenzuführen.

Probleme bei der Umsetzung und geringere Effekte als beim Training der Schülerinnen und Schüler durch externe Trainer hin (Dignath & Büttner, 2008). Trotz der Erkenntnis, dass die Vermittlung der Selbstregulationsstrategien durch die Lehrkräfte nachhaltiger wäre, hat nach Kenntnisstand der Autorin noch keine intensive Auseinandersetzung darüber stattgefunden, wie sich die neu zu implementierenden Inhalte in die bestehende Unterrichtspraxis einfügen.

So wäre es vor dem Hintergrund der in Abschnitt 1.1.1 dargestellten theoretischen Modellierung der allgemeinen Unterrichtsqualität denkbar, dass eine intensive, in die tägliche Unterrichtspraxis eingreifende Lehrerfortbildung durchaus eine sowohl positive als auch negative Veränderung anderer, nicht trainierter Unterrichtsaspekte bewirkt. Einen positiven Effekt könnte beispielsweise die Bereitstellung vorkonzipierter Unterrichtsstunden und Unterrichtsmaterialien haben. So könnte insbesondere der Unterricht von Lehrkräften, die aufgrund mangelnder Strukturierung Unterrichtszeit nicht effizient nutzen, von den bereitgestellten Stundenentwürfen profitieren. Auch die Empfehlungen zur selbstregulationsfördernden Bearbeitung von Aufgaben und die individualisierenden Elemente der Schülerförderung könnten zu positiven Transfereffekten auf nicht trainierte Aspekte der Schülerorientierung führen.

Mögliche negative Effekte könnten sich hingegen durch die Fokussierung der Lehrkraft auf die Selbstregulationsförderung ergeben. So nehmen Ophardt und Thiel (2008) in Anlehnung an die Befunde von Gruehn (1995) an, dass es für eine Lehrkraft zwar möglich ist, im Unterricht verschiedene Ziele zu erreichen (z.B. die gleichzeitige Förderung der Leistung und der Motivation). Dennoch weisen sie darauf hin, dass es in der Unterrichtspraxis aufgrund des eingeschränkten Zeitbudgets notwendig ist, eine Hierarchisierung von Unterrichtszielen vorzunehmen. Als Konsequenz dieser Aussage wäre es daher möglich, dass sich das Unterrichtsgefüge durch die fortbildungsbedingte Fokussierung auf die selbstregulationsspezifische Unterrichtsqualität verschiebt. Denkbar wäre demnach, dass die Integration der individualisierenden Elemente (z.B. Förderung der individuellen Bezugsnorm) der Schülerförderung im Fortbildungszeitraum sowohl in einer verstärkten Schülerorientierung als auch in einer weniger starken Konzentration auf die Klassenführung resultiert.

Ein weiteres Argument für eine fortbildungsbedingte Veränderung nicht-trainierter Aspekte der Unterrichtsqualität ergibt sich aus Forschungsbefunden zur Expertiseentwicklung von Lehrkräften. So wäre es in Anlehnung an Berliner (2001), Boshuizen (2004) und Bromme (1992) durchaus auch möglich, dass die Integration der Fortbildungsinhalte in die automatisierten Unterrichtsverläufe eingreift und so entwickelte Routinen unterbrochen

werden. Darauf, dass diese Unterbrechung selbst bei erfahrenen Lehrkräften eintreten kann, weisen Befunde von Borko und Putnam (1995) hin, die feststellen, dass eine interventionsbedingte Veränderung der Unterrichtspraxis auch Auswirkungen auf die Strategien der Klassenführung notwendig macht. Darüber hinaus berichtet Bromme (1992), dass Experten zur Lösung von für sie neuartigen Problemen und Aufgaben länger überlegen als unerfahrene Lehrkräfte.

In Anbetracht der großen Bedeutung der Unterrichtsqualität für das schulische Lernen (Scheerens & Bosker, 1997) wurden neben zahlreichen Studien zur Erfassung der Unterrichtsqualität in verschiedenen Schulformen (Baumert et al., 2004) auch Lehrerfortbildungen durchgeführt, die auf eine Verbesserung der Unterrichtsqualität abzielten. Sowohl die gezielte Förderung als auch die Evaluation der Fortbildungseffekte sind dabei jedoch häufig nur auf eine der drei Dimensionen wie bspw. die Klassenführung (z.B. Havers, 2010) oder die kognitive Aktivierung (z.B. Klieme et al., 2006) konzentriert. Es konnte daher zwar gezeigt werden, dass die drei Dimensionen der Unterrichtsqualität im Rahmen von Fortbildungen gezielt beeinflusst werden können, eine systematische Überprüfung der gleichzeitigen Wirkung von selbstregulationsfördernden Fortbildungen auf nicht trainierte Merkmale der Unterrichtsgestaltung steht jedoch noch aus.

Inwieweit sich durch eine mögliche Veränderung nicht trainierter Aspekte der Unterrichtsqualität auch Implikationen für die Annahmen zur Auswirkung von selbstregulationsfördernden Fortbildungen auf die Mathematikleistung ergeben, soll im nächsten Abschnitt erörtert werden.

1.2.4 Wirkweise von Lehrerfortbildungen zur Förderung der Selbstregulation auf die Mathematikleistung der Schülerinnen und Schüler

Hinsichtlich der Wirkweise von Lehrerfortbildungen zur Förderung der Selbstregulationsfähigkeit stellen Spörer und Glaser in ihrem Editorial fest, dass „oftmals zwingende Belege für die Annahme [fehlen], dass die Komponenten des selbstregulierten Lernens ein wirksames Element der Intervention sind“ (2010, S. 172). Sie thematisieren damit eine Problematik, die sich strenggenommen nicht nur auf die Wirkweise von Lehrerfortbildungen zur Förderung der Selbstregulation, sondern Interventionsstudien allgemein bezieht.

Dies spiegelt sich auch in der Feststellung von Dignath und Büttner (2008) wider, dass zahlreiche Interventionsstudien zur Förderung der Selbstregulation von Schülerinnen und Schülern ausschließlich eine Veränderung der Schülerleistung als Indikator für eine erfolgreiche Vermittlung der Trainingsinhalte heranziehen. Erst jüngere Studien untersuchen

in Mediationsanalysen, ob die Intervention tatsächlich zur postulierten Verbesserung der Lernstrategien der Schülerinnen und Schüler führt und ob diese dann wiederum einen Effekt auf die Lernleistung der Schülerinnen und Schüler hat (Brunstein & Glaser, 2011; Schünemann et al., 2013).

Diese Studien nehmen jedoch nur die Wirkmechanismen nach der Vermittlung der Trainingsinhalte an die Schülerinnen und Schüler in den Blick. Außerdem wurden die Trainingsinhalte von geschulten Assistenten anstelle von Lehrkräften vermittelt. Im Hinblick auf die stärkeren Effekte für durch externe Trainer anstatt durch fortgebildete Lehrkräfte durchgeführte Schülertrainings (Dignath & Büttner, 2008) muss jedoch angenommen werden, dass die Implementation in die Unterrichtspraxis eine nicht zu unterschätzenden Variable im Vermittlungsprozess ist. Es kann daher festgestellt werden, dass diese Studien zwar einen ebenfalls interessanten Ausschnitt im Wirkungsgefüge von selbstregulationsfördernden Interventionen untersuchen, nicht aber die Frage beantworten, wodurch die berichteten Effekte von Lehrerfortbildungen zur Förderung der Selbstregulation auf die Mathematikleistung der Schülerinnen und Schüler zu erklären sind.

So verweisen Spörer und Glaser (2010) in ihrem Editorial auch auf die Erkenntnis einiger Interventionsstudien, dass Lehrkräfte erlernte Fortbildungsinhalte häufig orientiert an den Gegebenheiten ihres Unterrichts modifizieren. So ist der Transfer von Trainingsinhalten in tägliche Unterrichtspraxis ein komplexer Prozess, dessen erfolgreicher Verlauf eine Mindestdauer, eine Kombination von Theorie und Praxis und die Bereitstellung von Unterrichtsmaterialien erfordert. In Anbetracht dieser hohen Anforderungen an Interventionen, die mithilfe von zuvor fortgebildeten Lehrkräften auf eine Verbesserung der Selbstregulation von Schülerinnen und Schülern zielen, stellt sich die Frage, ob frühere Fortbildungen die zuvor genannten Kriterien erfüllten und damit überhaupt einen Effekt auf die Unterrichtspraxis hatten. Außerdem stellt sich die Frage, ob Fortbildungen und der damit verbundene Eingriff in die Unterrichtspraxis nicht möglicherweise auch nicht intendierte Veränderungen der Basisdimensionen der Unterrichtsqualität herbeiführen, die die Schülerleistung beeinträchtigen.

Wie von Perels, Dignath und Schmitz (2009) angeführt, verwenden Lehrkräfte gemäß den Aussagen von Hamman, Berthelot, Saia und Crowley (2000) nur 9% der Redeanteile im Unterricht für die Behandlung von Selbstregulationsstrategien und nur 10% der Lehrkräfte im Primarbereich fördern die Anwendung von Selbstregulationsstrategien während der Bearbeitung von Aufgaben (Moely, Santulli & Obach, 1995). Sollte die Vermittlung solcher Strategien im Unterricht durch den Besuch einer Fortbildung nun einen deutlich größeren

Anteil einnehmen als zuvor, könnte dies durchaus ebenfalls einen Einfluss auf die Mathematikleistung haben. Insbesondere der Bereich der Klassenführung, der sich u.a. in einer Studie von Fauth, Decristan, Rieser, Klieme und Büttner (2014) als prädiktiv für die Mathematikleistung der Schülerinnen und Schüler erwiesen hat, könnte durch die Umsetzung von Stundenentwürfen und die stärkere Konzentration auf einen schülerorientierten Unterricht einen zusätzlichen Einfluss auf die Mathematikleistung aufweisen.

Basierend auf diesen Überlegungen könnte man annehmen, dass sich eine Lehrerfortbildung zur Förderung der Selbstregulation von Schülerinnen und Schülern nicht nur auf fokale Trainingsinhalte, sondern auch auf die allgemeine Unterrichtsqualität auswirkt. Ebenfalls durch die zuvor dargelegten Überlegungen begründet, könnte der Effekt der Fortbildung auf die Mathematikleistung über diesen möglichen Effekt auf untrainierte Aspekte der Unterrichtsqualität vermittelt werden anstatt—wie bisher angenommen—durch die Umsetzung der fokalen Trainingsinhalte. Eine empirische Überprüfung dieser Annahmen steht jedoch noch aus.

Im folgenden Abschnitt sollen aus den zuvor dargestellten theoretischen Grundlagen die dieser Arbeit zugrundeliegenden Forschungsfragen abgeleitet werden. Daran anschließend erfolgt eine kurze Vorstellung der Studien, welche die Klärung dieser zentralen Forschungsfragen zum Inhalt haben.

1.3 Fragestellungen der vorliegenden Arbeit

Wie zuvor ausgeführt wurden zum Thema Unterrichtsqualität eine Vielzahl an Studien durchgeführt, die sich bspw. mit der Dimensionalität von Unterrichtsqualität (Klieme et al., 2009; Pianta, La Paro & Hamre, 2008), deren Auswirkungen auf Schüleroutcomes wie Motivation und Leistungsentwicklung (Klieme et al., 2009; Seidel & Shavelson, 2007) oder auch mit schulformabhängigen Unterschieden der Unterrichtsqualität (Baumert et al., 2004; Kunter et al., 2005) beschäftigen. Auch die Frage, welche Instrumente zur Erfassung von Unterrichtsqualität aus forschungsmethodischer Sicht zu den validesten Aussagen führen, wurde bereits intensiv untersucht. Dabei wurde festgestellt, dass die Übereinstimmung zwischen Lehrer- und Schülerurteilen als niedrig bis moderat einzustufen ist. Diese Befunde zur Übereinstimmung von Lehrer- und Schülerurteilen wurden jedoch nach Kenntnis der Autorin nur unter Einbezug eines einzigen Messzeitpunkts ermittelt und es wurde bisher nicht untersucht, ob diese Befunde auch auf Längsschnittstudien übertragbar sind. Während jedoch Studien mit kleinen Stichprobengrößen auf eine Vielzahl von Datenquellen zurückgreifen können, sind Large-Scale-Studien mit sehr umfangreichen Stichprobengrößen auf den Einsatz

von Fragebögen angewiesen. In Anbetracht dieser Abhängigkeit von Fragebogendaten zur Ermittlung der Unterrichtsqualität in Large-Scale-Studien sind Erkenntnisse zur Übereineinstimmung von Lehrer- und Schülerurteilen in Längsschnitterhebungen von essentieller Bedeutung um sich für eine der beiden Datenquellen zu entscheiden. In Studie 1 wurde daher der bisher noch ungeklärten Frage nachgegangen, inwieweit sich Lehrer- und Schülerurteile vergleichbar zu Beobachterratings unter Einbezug mehrerer Messzeitpunkte in messzeitpunktspezifische und zeitüberdauernde Komponenten zerlegen lassen. Darüber hinaus untersucht Studie 1, ob diese aggregierten und um situationsspezifische Einflüsse bereinigten Lehrer- und Schülerurteile möglicherweise eine höhere Übereinstimmung zeigen als Lehrer- und Schülerurteile zu nur einem Messzeitpunkt.

Die zweite Fragestellung der vorliegenden Dissertationsschrift hat einen stärker inhaltlich ausgerichteten Fokus, indem sie den Schwerpunkt auf die Untersuchung der Auswirkungen von Lehrerfortbildungen auf die allgemeine Unterrichtsqualität richtet. Denn obwohl es Befunde zum Lernen von Lehrkräften (Berliner, 2001; Borko & Putnam, 1995; Boshuizen, 2004; Bromme, 1992) und zu Zielkonflikten in der Orchestrierung von Unterricht (Kunter, 2005; Oser & Baeriswyl, 2001) gibt, wurde bisher nicht der naheliegenden Frage nachgegangen, inwieweit sich die Einführung neuer Elemente in die Unterrichtspraxis im Rahmen von Lehrerfortbildungen auf die zuvor von der Lehrkraft praktizierte Unterrichtsgestaltung auswirkt. In Studie 2 soll daher der Frage nachgegangen werden, inwieweit sich die Lehrerfortbildung „Lernen mit Plan“ nicht nur auf die die fokalen, d.h. trainierten, Trainingsinhalte ausgewirkt hat, sondern sich auch eine Veränderung der allgemeinen Unterrichtsqualität und damit der nicht trainierten Unterrichtsaspekte ergibt.

Was aber bedeutet es, wenn sich im Rahmen einer Lehrerfortbildung, die eigentlich auf die Förderung der Selbstregulationskompetenzen der Schülerinnen und Schüler abzielt, auch untrainierte Aspekte der Unterrichtsqualität verändern? Wie im theoretischen Teil ausgeführt, wurde in zahlreichen Studien zur Förderung der Selbstregulation von Schülerinnen und Schülern auch eine Verbesserung der Mathematikleistung ermittelt. Diese positiven Effekte der Förderung der Selbstregulation bezogen sich dabei sowohl auf die Vermittlung von Selbstregulationsstrategien durch externe Trainer als auch durch zuvor in Lehrerfortbildungen geschulte Lehrkräfte. Sollte sich nun aber herausstellen, dass sich im Rahmen von Lehrerfortbildungen zur Förderung der Selbstregulation nicht nur ein Effekt auf die fokalen Trainingsinhalte ergibt, sondern auch auf die eigentlich nicht trainierten Aspekte der Unterrichtsqualität, so stellt sich die Frage, inwieweit die positiven Effekte auf die Mathematikleistung ausschließlich durch die Umsetzung der Fortbildungsinhalte

zurückzuführen sind. Diese Frage wird insbesondere dadurch untermauert, dass Dignath und Büttner (2008) in einer Metaanalyse zeigen konnten, dass die Effekte selbstregulationsfördernden Trainings stärker waren, wenn die Vermittlung der Trainingsinhalte von externen Trainer anstelle von Lehrkräften vorgenommen wurde. Studie 3 der vorliegenden Dissertation untersucht daher, inwieweit sich die Lehrerfortbildung „Lernen mit Plan“ auf die Mathematikleistung der Schülerinnen und Schüler auswirkt und inwieweit diese Effekte durch die Umsetzung der fokalen Trainingsinhalte oder aber durch die nicht-intendierte Veränderung der allgemeinen Unterrichtsqualität vermittelt werden.

Diese soeben ausgeführten Fragestellungen werden in drei Studien bearbeitet, deren Inhalte und forschungsmethodische Ansätze in den folgenden Abschnitten kurz dargestellt werden sollen, um dann in den Abschnitten 2 bis 4 der vorliegenden Dissertation in voller Länge präsentiert zu werden.

Gemeinsame Datengrundlage für alle drei Studien ist die Lehrerfortbildungsstudie „Lernen mit Plan“. Diese beinhaltet eine viertägige Lehrkräftefortbildung für Mathematiklehrkräfte der fünften Jahrgangsstufe an Baden-Württembergischen Haupt- und Werkrealschulen, die mithilfe eines Lehrermanuals geskriptete Unterrichtsstunden und Unterrichtsmaterialien bereitstellt. Zur Prüfung der Effekte von „Lernen mit Plan“ wurde ein präpost- follow-up-Design implementiert, wobei die follow-up-Erhebung sieben Wochen nach Ende der Umsetzung des Programms im Unterricht stattfand. Es erfolgte eine randomisierte Zuteilung der Lehrkräfte auf die Experimentalgruppe und die Wartekontrollgruppe.

Unterschiede zwischen den Studien ergeben sich dadurch, dass die Studien abhängig von der jeweiligen Fragestellung unterschiedliche Messzeitpunkte, Konstrukte und Perspektiven einbeziehen. So bezieht sich die eher psychometrisch ausgerichtete Studie 1 auf alle drei Messzeitpunkte und berücksichtigt sowohl die Lehrer- als auch die Schülerperspektive. Allerdings beschränkt sich die Studie auf drei Konstrukte (eines pro Dimension), auf die theoretisch durch die Lehrerfortbildung keine Effekte erwartet wurden. Darüber hinaus wurde im Rahmen der durchgeführten Analysen für mögliche Effekte der Fortbildung auf diese Konstrukte kontrolliert. In Studie 2 werden ebenfalls beide Perspektiven und alle drei Messzeitpunkte einbezogen. Allerdings beziehen sich die Analysen in dieser Studie auf eine Vielzahl an Konstrukten, die zuerst zusammengefasst zu Kompositen und später auf Ebene der Einzelskalen ausgewertet werden. In Studie 3 wurde nahezu die gleiche Konstruktauswahl wie in Studie 2 verwendet, allerdings erfolgt eine Beschränkung auf Schülerurteile und auf die Messzeitpunkte 1 und 3. Alle Studien berücksichtigen dabei die hierarchische Struktur der

Schülerdaten, indem mithilfe von Mehrebenenanalysen die Gesamtvarianz in Varianz innerhalb der Klasse und Varianz zwischen den Klassen unterteilt wird. Hinsichtlich der Stichprobengröße lässt sich ebenfalls festellen, dass diese zwischen den Studien in Abhängigkeit von den jeweiligen Konstrukten und Messzeitpunkten variiert, da je nach Analyseverfahren fehlende Werte zum Ausschluss von Personen führen. Detaillierte Informationen dazu finden sich in den folgenden Kurzbeschreibungen der Studien.

Studie 1 mit dem Titel *„Teacher-Student Ratings of Instructional Quality: Decomposing Overall Agreement and Occasion-Specific Effects“* untersucht inwieweit sich durch Aggregation mehrerer Messzeitpunkte mithilfe von Faktorenanalysen messzeitpunkt-spezifische und messzeitpunktübergreifende Varianzanteile separieren lassen. Anlass zu dieser Studie gab einerseits die Diskrepanz zwischen der großen Bedeutung von Unterrichtsqualität zur Vorhersage von wichtigen Schüleroutcomes wie der Leistungsentwicklung und der gleichzeitig (akzeptierten) geringen Übereinstimmung der Maße zur Erfassung der Unterrichtsqualität wie Lehrer- und Schülerratings. Andererseits haben bereits einige Studien zur Unterrichtsqualität auf mögliche Einflüsse auf Unterrichtsratings durch die leichte Variabilität des Unterrichts und durch messzeitpunktspezifische Wahrnehmungstendenzen hingewiesen. Bisher steht eine empirische Untersuchung der Auswirkungen der messzeitpunktspezifischen Wahrnehmungs-tendenzen auf die Übereinstimmung zwischen den Lehrer- und Schülerurteilen jedoch noch aus. Wie in Abbildung 2 dargestellt, untersucht Studie 1 daher die Höhe des Anteils situationsspezifischer Varianz in Lehrer- und Schülerratings und damit verbundene Konsequenzen für die Übereinstimmung beider Perspektiven. Die Mehrebenenfakto-renanalysen der Studie beziehen sich dabei auf Lehrer- und Schülerurteile der drei Konstrukte Klassenführung, Zielsetzung und Autonomieunterstützung, die über einen Zeitraum von drei Monaten bei 74 Lehrkräften und deren Schülerinnen und Schülern erhoben wurden.

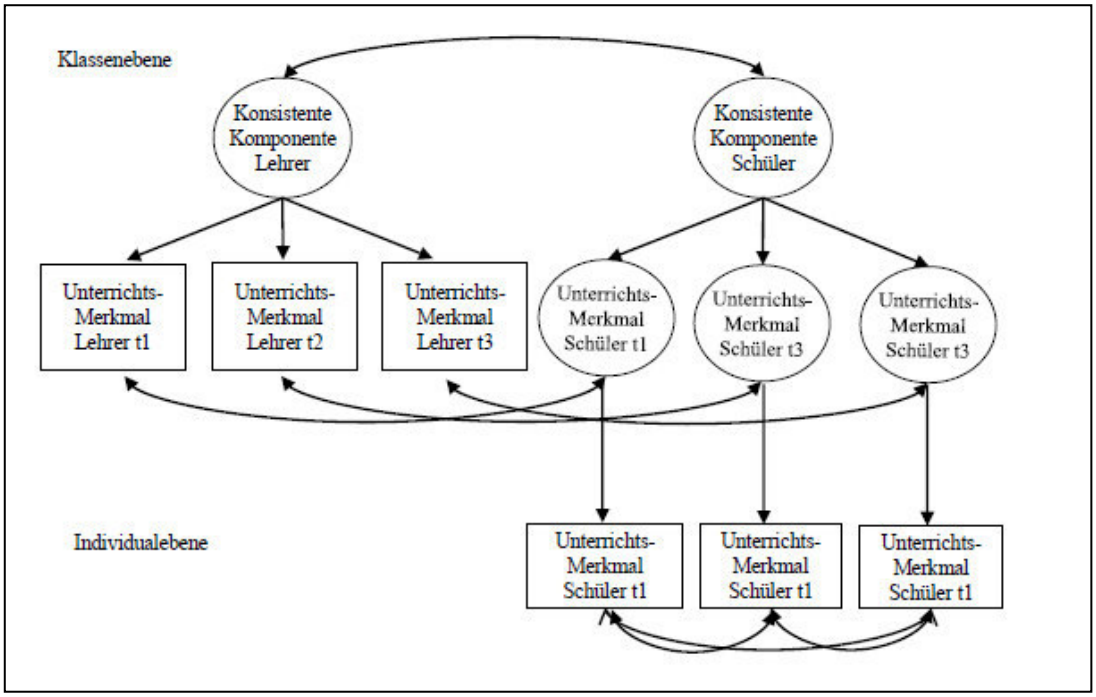

Abb. 2: Analysemodell von Studie 1.

Studie 2 mit dem Titel *„Förderung des selbstregulierten Lernens durch die Lehrkräftefortbildung «Lernen mit Plan»: Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität“* untersucht unter Bezugnahme auf fragebogenbasierte Lehrer- und Schülerurteile von allen drei Messzeitpunkten die Forschungsfrage, ob das Training dazu führt, dass die selbstregulationsspezifische Unterrichtsqualität gesteigert wird. Die zu überprüfenden Hypothesen sind dabei, dass das Training in der Experimentalgruppe im Vergleich zur Kontrollgruppe sowohl aus Lehrer- als auch aus Schülersicht zu einer unmittelbaren und langfristigen verstärkten Selbstregulationsförderung im Unterricht führte. Zweitens wird untersucht, welche Effekte die Fortbildungsteilnahme im Hinblick auf Indikatoren der allgemeinen Unterrichtsqualität hatte. Hinsichtlich der Auswirkungen auf die allgemeine Unterrichtqualität wird in Anbetracht von potenziell konfligierenden Konsequenzen und aufgrund des Mangels an entsprechenden Forschungsbefunden zwar eine Veränderung der allgemeinen Unterrichtsqualität erwartet, es werden jedoch keine gerichteten Vorerwartungen bzgl. der unmittelbaren und langfristigen Entwicklung formuliert. So ist einerseits denkbar, dass bspw. die individualisierenden Elemente der Schülerförderung zu positiven Transfereffekten auf weitere Aspekte der Schülerorientierung führen. Auch ist denkbar, dass die starke Nutzung von manualbasierten Stundenentwürfen und vorgeschlagenen Fördermaßnahmen zu einer besseren Strukturierung des Unterrichts sowie

verstärkten Nutzung kognitiv aktivierender Elemente führt. Aber auch negative Effekte durch die Unterbrechung eingespielter Unterrichtsroutinen werden für möglich gehalten. Das zu untersuchende Wirkmodell kann Abbildung 3 entnommen werden. Für die Analysen wurden Lehrer- und Schülerratings von insgesamt 75 Lehrkräften (Experimentalgruppe: 47 Lehrkräfte; Wartekontrollgruppe: 28 Lehrkräfte) herangezogen. Basierend auf Globalmaßen und später auf Ebene der einzelnen Skalen werden mithilfe von Differenzwerten Unterschiede zwischen den Effekten der Fortbildungsteilnahme sowohl auf die Experimental- und die Kontrollgruppe als auch auf fokale und nicht-trainierte Aspekte der Unterrichtsgestaltung ermittelt.

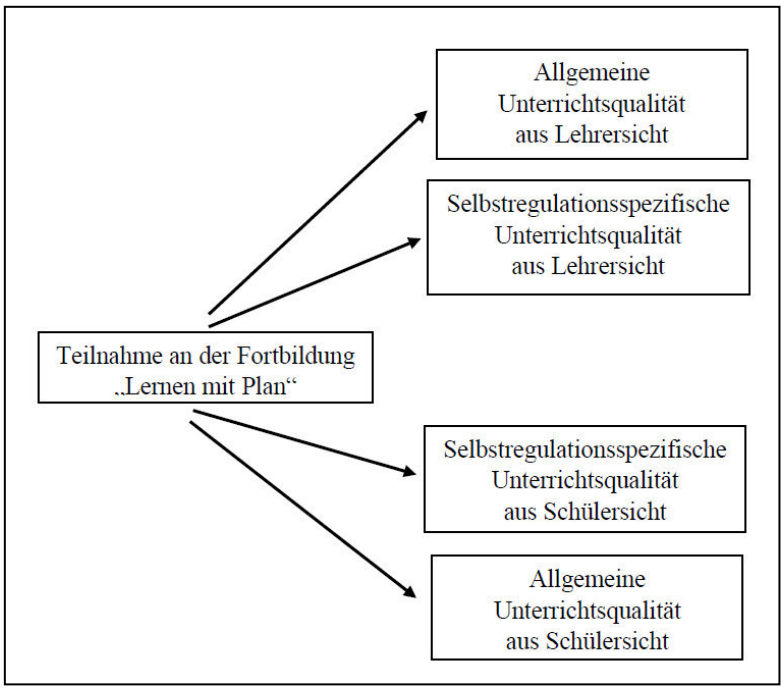

Abb. 3: Arbeitsmodell zur Wirkungsweise der Fortbildung „Lernen mit Plan“ auf die selbstregulations-spezifische und allgemeine Unterrichtsqualität aus Lehrer- und Schülersicht

Studie 3 mit dem Titel “Opening the Black Box: Are Effects of a Teacher Training to Foster Students-Self-Regulation on Students’ Math Competencies Mediated by Trained or Untrained Aspects of Teaching Practice?“ untersucht wie in Abbildung 4 dargestellt, ob die häufig durch Selbstregulationstrainings erzielten positive Effekte auf die Mathematikleistung von Schülerinnen und Schülern ausschließlich durch eine Umsetzung der Selbstregulationsförderung im Unterricht vermittelt wird oder ob hierbei auch Effekte möglicher Veränderungen der allgemeinen Unterrichtsqualität auf die Mathematikleistung identifizierbar

sind. Hierbei wird angeknüpft an die Befunde aus Studie 2, in der untersucht wird, ob die selbstregulationsfördernde Lehrerfortbildung „Lernen mit Plan“ auch einen Effekt auf die untrainierte, allgemeine Unterrichtsqualität hat. So legt der empirisch vielfach belegte Zusammenhang zwischen Aspekten der allgemeinen Unterrichtqualität und der Leistungsentwicklung von Schülerinnen und Schülern nahe, dass eine Veränderung der allgemeinen Unterrichtsqualität im Rahmen der Lehrerfortbildung „Lernen mit Plan“ auch einen Effekt auf die Mathematikleistung zum dritten Messezeitpunkt haben könnte. Basierend auf den Schülerurteilen von 686 Schülerinnen und Schülern aus 74 Klassen wurden unter Kontrolle der Prätestwerte in Mehrebenen-Mediationsanalysen ermittelt, ob ein Effekt der Fortbildung auf die durch standardisierte Mathematiktests ermittelte Mathematikleistung der Schülerinnen und Schüler zum dritten Messzeitpunkt sowohl durch die Umsetzung der Trainingsinhalte als auch durch die Veränderung der allgemeinen Unterrichtsqualität vermittelt wird.

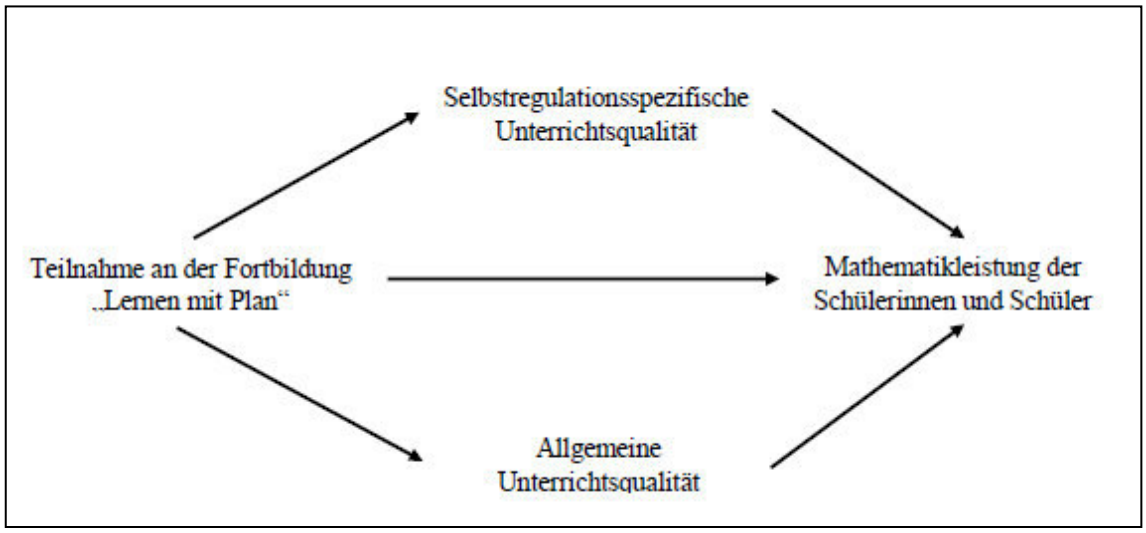

Abb. 4: Arbeitsmodell zur Wirkungsweise der Fortbildung „Lernen mit Plan“ auf die selbstregulationsspezifische und allgemeine Unterrichtsqualität und die Mathematikleistung der Schülerinnen und Schüler

1.4 Literatur

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153-166.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr. & Traver, H. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*(2), 341-358. doi: 10.1111/j.1744-6570.1997.tb00911.x

Baumert, J., Blum, W., Brunner, M., Krauss, S., Kunter, M. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. [Mathematics teaching from the perspective of the PISA students and their teachers]. In M. Prenzel, J. Baumert, W. Blum, R. Lerhmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (314-354). Münster: Waxmann-Verlag.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463-482. doi: http://dx.doi.org/10.1016/S0883-0355(02)00004-6

Borko, H. (2004). Professional Development and Teacher Learning: Mapping the Terrain. *Educational Researcher, 33*(8), 3-15. doi: 10.3102/0013189x033008003

Borko, H. & Putnam, R. T. (1995). Expanding a teachers' knowledge base: A cognitive psychologica perspective on professional development. In T. R. Guskey & M. Huberman (Hrsg.), *Professional development in education: New paradigms and practices* (35–66). New York: Teachers College Press.

Boshuizen, H. P. A. (2004). Does practice make perfect? A slow and discontinuous process. In H. P. A. Boshuizen, R. Bromme & H. Gruber (Hrsg.), *Professional learning: Gaps and transitions on the way from novice to expert* (3-8). Dordrecht: Kluwer Academic Press.

Bromme, R. (1992). *Der Lehrer als Experte : zur Psychologie des professionellen Wissens*. Bern {[u.a.]: Huber.

Brunstein, J. C. & Glaser, C. (2011). Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology, 103*(4), 922-938. doi: 10.1037/a0024622

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt - und Kriteriumsvalidität. [Quality of instruction: A matter of perspective?]*. Münster, Westfalen u.a.: Waxmann.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N. & Orphanos, S. (2009). Professional learning in the learning profession. A status report on teacher development in the United States and abroad. Dallas, TX: National Staff Development Council.

de Corte, E., Verschaffel, L. & van De Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. *British Journal of Educational Psychology, 71*, 531–559.

De Jong, R. & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51-85.

Desimone, L., Smith, T. & Frisvold, D. (2009). Survey measures of classroom Instruction: Comparing student and teacher reports. *Educational Policy*. doi: 10.1177/0895904808330173

Dignath, C. & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231-264. doi: 10.1007/s11409-008-9029-x

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*(0), 1-9. doi: http://dx.doi.org/10.1016/j.learninstruc.2013.07.001

Fullan, M. G. & Stiegelbauer, S. (1991). *The Meaning of Educational Change.* New York: Teachers College Press, Columbia University.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K. et al. (2011). Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gentry, M., Gable, R. K. & K., R. M. (2002). Students' perceptions of classrooms activities: Are there grade level and gender differences? . *Journal of Educational Psychology, 94*(3), 539-544.

Goldschmidt, P. & Phelps, G. (2007). Does teacher professional development affect content and pedagogical knowledge: How much and for how long? Los Angeles:

National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education and Information Studies, University of California.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *Am Psychol, 52*(11), 1182-1186.

Gruehn, S. (1995). Vereinbarkeit kognitiver und nichtkognitiver Ziele im Unterricht. *Zeitschrift für Pädagogik, 41*(4), 531-553.

Hamman, D., Berthelot, J., Saia, J. & Crowley, E. (2000). Teachers' coaching of learning and its relation to students' strategic learning. *Journal of Educational Psychology, 92*(2), 342-348.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* NY: Routledge.

Havers, N. (2010). Lässt sich effiziente Klassenführung lehren? Das Potenzial der Lehrertrainings. In J. Abel & G. Faust (Hrsg.), *Wirkt Lehrerbildung? Antworten aus der empirischen Forschung* (283-290). Münster: Waxmann.

Kane, T. J. & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. In B. M. G. Foundation (Hrsg.), *MET Project Research Paper*.

Kirkpatrick, D. (1996). Great Ideas Revisited. *Training & Development, 50*(1), 6.

Klieme, E., Lipowsky, F., Rakoczy, K. & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". [Quality dimensions and effectiveness of mathematics instruction. Theoretical background and selected findings of the Pythagoras project]. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule.* (127-146). Münster, Westfalen u.a.: Waxmann.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom.* (137-160). Münster u.a.: Waxmann.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung. In B. f. B. u. Forschung (Hrsg.), *TIMMS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (43-58). Bonn: BmBF.

Klug, J., Ogrin, S., Keller, S., Ihringer, A. & Schmitz, B. (2011). A plea for self-regulated learning as a process: Modelling, measuring and intervening. *Psychological Test and Assessment Modeling, 53*(1), 51-72.

Komorek, E., Bruder, R., Collet, C. & Schmitz, B. (2007). Contents and results of an intervention in maths lessons in secondary level I with a teaching concept to support mathematic problem-solving and self-regulative competencies. In M. Prenzel (Hrsg.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (175-196). Münster: Waxmann.

Koziol, S. M. & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research, 79*(4).

Krammer, K., Ratzka, N., Klieme, E., Pauli, C. & Reusser, K. (2006). Learning with classroom videos: Conception and first results of an online teacher-training program. *Zeitschrift für Didaktik der Mathematik, 38*, 422-432.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231-251. doi: 10.1007/s10984-006-9015-7

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W. et al. (2005). Der Mathematikunterricht der PISA- Schülerinnen und Schüler. *Zeitschrift für Erziehungswissenschaft, 8*(4), 502-520. doi: 10.1007/s11618-005-0156-8

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (85-113). Münster: Waxmann.

Labuhn, A. S., Zimmerman, B. J. & Hasselhorn, M. (2010). Enhancing students‘ self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), 173-194.

Lanahan, L., McGrath, D. J., McLaughlin, M., Burian-Fitzgerald, M. A. & Salganik, L. (2005). Fundamental problems in the measurement of instructional processes: Estimating reasonable effect sizes and conceptualizing what is important to measure. Washington, DC: American Institutes for Research.

Lievens, F., Reeve, C. L. & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*(6), 1672-1682. doi: 10.1037/0021-9010.92.6.1672

Lipowsky, F. (2010). Lernen im Beruf - Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In F. Müller, A. Eichenberger, M. Lüders & J. Mayr (Hrsg.), *Lehrerinnen und Lehrer lernen - Konzepte und Befunde zur Lehrerfortbildung* (51-72). Münster: Waxmann.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527-537. doi: 10.1016/j.learninstruc.2008.11.001

Marsh, H. W. & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(2), 116-145. doi: 10.1080/10705519409539968

Moely, B. E., Santulli, K. A. & Obach, M. S. (1995). Strategy Instruction, Metacognition, and Motivation in the Elementary School Classroom. In F. E. Weinert & W. Schneider (Hrsg.), *Memory performance and competencies – Issues in growth and development* (301-321). Mahwah, NJ: Erlbaum.

Mokhlesgerami, J., Souvignier, E., Rühl, K. & Gold, A. (2007). Naher und weiter Transfer eines Unterrichtsprogramms zur Förderung der Lesekompetenz in der Sekundarstufe I. *Zeitschrift für Pädagogische Psychologie, 21*(2), 169-180.

Ophardt, D. & Thiel, F. (2008). Klassenmanagement als Basisdimension der Unterrichtsqualität. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (2. vollst. überarbeitete Aufl., 259-282). Wiesbaden: VS Verlag für Sozialwissenschaften.

Oser, F. K. & Baeriswyl, F. J. (2001). Choreographies of Teaching: Bridging Instruction to Learning. In V. Richardson (Hrsg.), *Handbook of Research on Teaching* (4 Aufl.). Washington: American Educational Research Assotiation.

Otto, B. (2007). *SELVES. Schüler-, Eltern-, und Lehrertraining zur Vermittlung effektiver Selbstregulation.* (Dissertation), Logos Verlag Berlin GmbH, Darmstadt.

Perels, F., Dignath, C. & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in

regular math classes. *European Journal of Psychology of Education, 24*(1), 17-31. doi: 10.1007/BF03173472

Perels, F., Gürtler, T. & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction, 15*(2), 123-139. doi: 10.1016/j.learninstruc.2005.04.010

Pianta, R. C., La Paro, K. M. & Hamre, B. K. (Hrsg.). (2008): *Classroom Assessment Scoring System (CLASS).* (Ausgabe Ausg. Heft Heft). Baltimore: Paul H. Brookes.

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher, 31*(7), 3-14.

Rosenshine, B. (1970). Evaluation of Classroom Instruction. *Review of Educational Research, 40*(2), 279-300. doi: 10.3102/00346543040002279

Rozendaal, J. S., Minnaert, A. & Boekaerts, M. (2005). The influence of teacher perceived administration of self-regulated learning on students' motivation and information-processing. *Learning and Instruction, 15*(2), 141-160. doi: 10.1016/j.learninstruc.2005.04.011

Scheerens, J. & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Schünemann, N., Spörer, N. & Brunstein, J. C. (2013). Integrating self-regulation in whole-class reciprocal teaching: A moderator–mediator analysis of incremental effects on fifth graders' reading comprehension. *Contemporary Educational Psychology, 38*(4), 289-305. doi: http://dx.doi.org/10.1016/j.cedpsych.2013.06.002

Seidel, T., Rimmele, R. & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction, 15*(6), 539-556. doi: 10.1016/j.learninstruc.2005.08.004

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). doi: 10.3102/0034654307310317

Spörer, N. & Glaser, C. (2010). Förderung selbstregulierten Lernens im schulischen Kontext. *Zeitschrift für Pädagogische Psychologie, 24*(3), 171-175. doi: 10.1024/1010-0652/a000014

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz.* Retrieved from http://kola.opus.hbz-nrw.de/volltexte/2008/234

Waldis, M., Grob, U., Pauli, C. & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität - Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (171-208). Münster: Waxmann.

Weinstein, R. S. (1985). Student mediation of classroom expectancy effects. In J. B. Dusek (Hrsg.), *Teacher expectancies.* Hillsdale, NJ: Erlbaum.

Wubbels, T., Brekelmans, M. & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*(1), 47-58.

Yoon, K. S., Duncan, T., Lee, R. W.-Y., Scarloss, B. & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement *Issues & Answers Report*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Hrsg.), *Self-regulated learning and academic achievement: Theoretical perspectives* (1–37). Mahwah, NJ: Erlbaum.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183.

Zimmerman, B. J. & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal, 31*(4), 845-862.

2

Teacher-Student Ratings of Instructional Quality: Decomposing Overall Agreement and Occasion-Specific Effects

Werth, S., Wagner, W., Trautwein, U., Göllner, R., Voss, T. & Schmitz, B. (eingereicht). Teacher-Student Ratings of Instructional Quality: Decomposing Overall Agreement and Occasion-Specific Effects. *Journal of Educational Psychology.*

Abstract

Prior research has shown that both teacher and student ratings of instructional quality predict important outcome variables such as achievement, motivation, and well-being. Yet research has also shown that the correlation between teacher and student ratings of instructional quality is, at best, moderate. Explanations for this low to moderate association include the assumption that teacher and student ratings are influenced by a perspective-specific referent period. The present study thus investigated whether teacher and student ratings of mathematics lessons in the fifth grade also consist of stable and time-specific components. Based on the idea that both instructional variability and the specific referent period influence the agreement between teacher and student ratings, the study further investigated whether teacher-student agreement would increase when aggregated measures were used instead of single-wave measurements. Three waves of teacher and student ratings of instructional quality (classroom management, goal clarity, support of autonomy) were gathered across a time period of 3 months from a total of 74 classes. Results from multi-level factor analyses indicate that teacher and student ratings of instructional quality include both stable and variable components. Furthermore, the correlation between teacher and student ratings of classroom management and goal clarity increased when using aggregated measures of instructional quality, whereas the association between teacher and student ratings of the support of autonomy, a construct with a stronger reference to the individual student, did not increase. The authors conclude that aggregated measures might be a way to overcome the influence of perspective-specific referent periods.

Keywords: teacher-student ratings of instructional quality, instructional variability, repeated measurements

Teacher-Student Ratings of Instructional Quality: Decomposing Overall Agreement and Occasion-Specific Effects

Instructional quality comprises all teacher-student interactions that stimulate students' cognitive and motivational development (Caroll, 1963; Weinert, Schrader, & Helmke, 1989) and is crucial to students' engagement and learning (Hattie, 2008; Scheerens & Bosker, 1997; Seidel & Shavelson, 2007). To a much larger extent than students' individual backgrounds, instructional quality is modifiable (Lanahan, McGrath, McLaughlin, Burian-Fitzgerald, & Salganik, 2005) and has thus been targeted by many large-scale studies (Kane & Staiger, 2012; Mullis et al., 1997, 1998; Organisation for Economic Co-operation and Development [OECD], 1999) and interventions (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Garet et al., 2010) as a way to improve educational systems.

In order to gain a deeper understanding of the effects of instructional quality, it is essential to use instruments with high reliability and validity. In this context, the low to moderate association between teacher and student surveys—two prominent measures of instructional quality—has presented a continuing challenge in instructional quality research (Clausen, 2002; Desimone, Smith, & Frisvold, 2009; Fauth, Decristan, Rieser, Klieme, & Büttner, in press; Kunter & Baumert, 2006).

For many years, this low to moderate association has been explained as the consequence of the idiosyncratic validity of each perspective (Clausen, 2002; De Jong & Westerhof, 2001; Desimone et al., 2009; Kunter & Baumert, 2006). However, it can also be argued that the low to moderate correlation between the teacher and student perspectives is a consequence of the low retest reliability of teacher and student ratings due to the volatility of classroom instruction and perspective-specific referent periods (Clausen, 2002; Lanahan et al., 2005; Weinstein, 1985). In fact, recent studies have shown that observer ratings—a third prominent measure of instructional quality—become more representative of *general* instructional quality when several measurement occasions are aggregated (Kane & Staiger, 2012; Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014). Findings from this research raise the question of whether the reliability of teacher and student ratings (i.e., the extent to which they reflect *general* instructional quality instead of occasion-specific [perception of] teaching practice) can also be increased when several measurement occasions are aggregated. The application of aggregated ratings would help to separate the amount of variance due to a consistent teaching practice component (which might reflect general teaching practice or at least a consistent perception of teaching practice) from variance due to a variable teaching practice component (which either

reflects a variable perception of instructional quality or actual variable teaching practice due to different topics within a day, a week, and a school year; Curby et al., 2011; Hamre, Pianta, & Chomat-Mooney, 2009; Kane & Staiger, 2012; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Praetorius et al., 2014; Seidel & Prenzel, 2006). Assuming that instructional variability in combination with differing referent periods in teacher and student ratings might be responsible for low teacher-student agreement, separating the consistent and variable components in teacher and student ratings might lead to greater agreement between these measures.

The present study used data from a study of 74 fifth-grade classrooms in which instructional quality was measured three times across a period of 3 months. We examined the extent to which the three different measurement occasions in our study reflected the consistent and variable components in teacher and student ratings of instructional quality. Based on the assumption that aggregated measures of instructional quality can help to overcome the influence of instructional variability and rater-specific referent periods on teacher and student ratings, we further investigated whether teacher-student agreement would increase when aggregated measures of the respective instructional aspect were used.

Domains of Instructional Quality

Whereas there are many different theoretical and empirical approaches for classifying components of instructional quality, there is some consensus with regard to three global domains (Baumert et al., 2004; Klieme, Pauli, & Reusser, 2009; Pianta, La Paro, & Hamre, 2008): class organization, instructional support, and emotional support. According to this classification, class organization comprises aspects such as classroom management and time on task (Klieme et al., 2009). Instructional support covers teaching components such as cognitive stimulation, goal clarity, and task-contingent feedback styles and relates to any instructional practice that fosters students' higher level thinking (Klieme et al., 2009; Seidel, Rimmele, & Prenzel, 2005). Emotional support covers observable features of teacher-learner interactions such as support of autonomy, participation, and teacher support (Klieme et al., 2009). Whereas the first two domains have been shown to primarily affect students' cognitive growth and learning processes (Klieme, Lipowsky, Rakoczy, & Ratzka, 2006; Klieme et al., 2009), research by Klieme et al. (2006; 2009) and Lipowsky et al. (2009) has indicated that emotional support has a direct effect on students' motivational development and an indirect effect on students' achievement.

Relative Agreement between Teacher and Student Ratings of Instructional Quality

Prior research findings have shown that both teacher and student ratings have face validity (Kunter & Baumert, 2006) and are predictive of student outcomes (Clausen, 2002; De Jong & Westerhof, 2001; Desimone et al., 2009; Kunter & Baumert, 2006). However, research on instructional quality has found a surprisingly low association (i.e., relative agreement) between teacher and student ratings of instructional quality, thus indicating low convergent validity (Clausen, 2002; Desimone et al., 2009; Fauth et al., in press; Kunter & Baumert, 2006). Clausen (2002), for instance, reported relative agreement between teacher and student ratings of the same construct that ranged from $-.28 \leq r \leq .42$. Kunter and Baumert (2006) compared teacher and student ratings of instructional quality and found correlations among latent factors that ranged from $.09 \leq r \leq .64$.

The construct-specific degree of relative agreement between the two perspectives may have at least four different reasons. First, the degree of inference may help to explain the strength of the association between ratings from different perspectives. According to Rosenshine (1970), ratings of classroom instruction can be classified as *low-inference* ratings of "specific, denotable, relatively objective behaviors" (p. 281) and *high-inference* ratings that "lack such specifity" (p. 281). Whereas the highest agreement ($r = .64$ in Kunter & Baumert, 2006; $r = .45$ in Fauth et al., in press) was found for low-inference ratings such as (*inefficient*) *classroom management*, the lowest agreement (ranging from $.09 \leq r \leq .24$) was found for constructs such as *types of tasks* and *support of cognitive autonomy*, which can be regarded as high-inference items.

Second, the specific item wording can be assumed to be a relevant predictor of the association between constructs assessed by different perspectives. The use of similar or parallel item wordings, however, does not guarantee that theoretically identical constructs are assessed (Kunter & Baumert, 2006). Urdan (2004), for instance, found that teachers' as well as students' interpretations of aspects of instructional quality may be quite different from the theoretically defined constructs. Therefore, in order to assess a given construct from different perspectives, the use of identical wordings may not be the preferred approach (Fauth et al., in press).

Third, different dimensions of instructional quality (even within the same domain) may be characterized as more or less content-dependent (Praetorius et al., 2014). Some constructs may be more or less suited for different kinds of lessons: Cognitive activation, for instance, may be most relevant in introduction lessons (Praetorius et al., 2014). A study by Tsai et al.

(2008) found a large variability regarding the perceived support of autonomy, which may also be due, at least in part, to changing contents of the lessons. Classroom management, in contrast, was shown to be rather stable across lessons compared to other constructs for observer ratings (Kane & Staiger, 2012; Praetorius et al., 2014) and student ratings (Rakoczy, 2008). If students (or teachers) rate instructional quality regarding long referent periods, the rating may strongly depend on the specific lessons (with a specific content) the students have in mind. This, in turn, should have a stronger impact on content-dependent constructs.

Fourth, teacher ratings of their own instructional quality may be influenced by self-serving strategies (Clausen, 2002; Fauth et al., in press; Kunter & Baumert, 2006; Wubbels, Brekelmans, & Hooymayers, 1992). The impact of this bias may differ from construct to construct and could vary across teachers – and, thereby, lower relative agreement between teacher and observer or student ratings. Fauth et al. (in press), for instance, argue that constructs belonging to the domain class organization may be only slightly biased by self-serving strategies, which – besides the low degree of inference of such indicators – may help explain the strong association between student and teacher ratings found in their study for this domain.

Whereas some researchers have explained the gap between the two perspectives by different sources of rater bias and have tried to reduce it by controlling for teachers' and students' individual background (Desimone et al., 2009), other researchers have concluded that there is probably no single "true" instructional quality but rather that each perspective has idiosyncratic validity (De Jong & Westerhof, 2001; Kunter & Baumert, 2006). Given the undisputed importance of instructional quality (Cornelius-White, 2007; Hattie, 2008; Scheerens & Bosker, 1997; Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993), the low agreement between two of the most often applied measures of instructional quality poses a serious challenge for research. The limited convergent validity of the measures can lead to incorrect conclusions about the influence of instructional quality on learning outcomes. Furthermore, the measurement of the effects of interventions and reforms of instructional quality might be impeded as well.

Consistent and Time-Specific Aspects of Teaching Practice in Ratings of Instructional Quality

Most reports of low associations between teacher and student ratings of instructional quality have been drawn from cross-sectional studies that are unable to take into account variability in (perceived) instructional practice. Yet, recent classroom-observation studies

have shown that instructional patterns vary considerably between lessons taught by the same teacher (Curby et al., 2011; Hamre et al., 2009; Kane & Staiger, 2012; Malmberg et al., 2010; Seidel & Prenzel, 2006). According to Curby et al. (2011), characteristics of a teacher's practice can vary from lesson to lesson, comprising on the one hand consistencies across their teaching practice and, on the other hand, variable features of classroom interaction that depend on, for instance, the topic of the lesson or the group of students (Kane & Staiger, 2012; Malmberg et al., 2010; Seidel & Prenzel, 2006). When measuring instructional quality, this instructional variability has to be taken into account, for instance, by measuring classroom quality several times instead of relying on single-wave data (Kane & Staiger, 2012; Praetorius et al., 2014; Seidel & Prenzel, 2006). In most studies, however, the assessment of instructional quality by teacher and student ratings occurs only once rather than several times. The question thus arises as to whether aggregating several measurement occasions could lead to higher teacher-student agreement due to a larger agreement between the two perspectives concerning the consistent teaching practice component of ratings of instructional quality rather than concerning the variable teaching practice component.

Although the influence of instructional variability on teacher and student ratings and thus on the agreement between the two perspectives has not yet been investigated empirically, Clausen (2002) suggested that taking repeated measurements could help to reduce the influence of the inherent referent period on ratings of instructional quality. Referring to Weinstein (1985), he assumed that students might be strongly influenced by impressive situations from past days or weeks when asked about "typical" instructional quality. This suggestion is in line with Lanahan et al. (2005, p. 1), who assumed that items that require teachers and students to recall far back in time (e.g., items that ask for typical instruction over the past semester or year) will probably not be answered reliably. Due to instructional variability, however, items that ask teachers and students about the last lesson might also be not very representative of the teachers' (average) practice over the academic year.

Findings from the *Measures of Effective Teaching* (MET) project that used classroom observation in order to measure instructional quality (Kane & Staiger, 2012) have indicated that measurements based on several measurement occasions can help to better capture the average of the respective instructional feature (see also Praetorius et al., 2014). This could lead to more reliable measures (i.e., measures that reflect the "true score" or consistent teaching practice to a higher degree than single-wave measurement). Kane and Staiger (2012) showed that 23% to 32% of the variance in observer ratings of instructional quality could be attributed to an individual teacher's consistent teaching pattern. When averaging the scores

from the single lessons (rated by different raters), the reliability of the measures in the MET project increased. Instead of formerly 23% to 32%, the consistent teaching practice component then accounted for about 63% (class organization: 62%; instructional support: 55%; emotional support: 61%). This implies that some observation-to-observation variability has actually to be attributed to differences between the observed lessons (in the MET project 18% to 28%) and to differences between the raters (10% to 14%) and not to consistent differences between different teachers' teaching practice. Based on these findings, Kane and Staiger (2012, p. 36) concluded "Therefore, to capture persistent differences in teacher practice, it is important to average across more than one lesson".

However, the extent to which the findings from the studies of Kane and Staiger (2012) and Praetorius et al. (2014) would also apply to student or teacher ratings of instructional quality is not clear. Teacher and student ratings differ from observer ratings in several aspects: In contrast to observers, teachers and students are typically asked to apply a longer referent period instead of referring only to a single lesson. Furthermore, it seems possible that teacher and student ratings of instructional quality might additionally be affected by the use of repeated measurements to a much higher extent than observer ratings. For instance, in contrast to observers teachers and students are usually not familiar with the applied instruments like tests or questionnaires (Lievens, Reeve, & Heggestad, 2007) and are not trained to evaluate a teacher’s performance. Because teachers and students usually have not had to rate the teacher’s performance with respect to the applied items before, it has to be considered that their ratings at the first measurement occasion might be more intuitive and thus more strongly influenced by impressions from the last lesson or last week. In case that teachers and students are indeed able to recognize a consistent teaching component when asked about instruction several times, ratings from a second or third measurement occasion might reflect such a consistent component to a larger extent because teachers and students might become more familiar with the instrument and more sensitive to instructional quality in the respective classroom over time (Koziol Jr & Burns, 1986; Lievens et al., 2007). Furthermore, Lanahan et al. (2005) noted that teacher and student surveys are often not answered validly due to the decreasing commitment that results from the high burden that is linked to filling out long questionnaires. With regard to repeated measurement this would imply that due to a decrease in this commitment, the amount of residual variance in teacher and student ratings that cannot be attributed to a consistent teaching style might increase, whereas the variance due to the consistent teaching practice component would decrease.

Despite these differences, it seems reasonable to assume that also teacher and student ratings of instructional quality include a consistent teaching practice component which might reflect general teaching practice or at least a consistent perception of teaching practice as well as a variable teaching practice component, the latter reflecting impressions that are specific to the respective measurement occasion (e.g., actual variable teaching practice due to different topics within a day, a week, and a school year; Curby et al., 2011; Hamre et al., 2009; Kane & Staiger, 2012;Malmberg et al., 2010; Seidel & Prenzel, 2006). If this indeed is the case, the aggregation of teacher and student ratings from several measurement occasions would allow for extracting the consistent teaching practice component and thus potentially provide more reliable and valid measures of consistent instructional quality than single-wave measurements that also include the variable, occasion-specific component. Additionally, based on Clausen's (2002), Lanahan's (2005), and Weinstein's (1985) assumption that teachers and students apply different referent periods when they rate instructional quality (which would then be captured by the variable component and separated from the consistent component), the agreement between teacher and student ratings of the same construct might increase when teacher–student agreement is estimated for aggregated measures instead of single measurement occasions. With regard to student ratings, the aggregation of individual ratings to a class level measure already (largely) eliminates individual occasion-specific components. It is still possible, however, that even the aggregated measures at the class level reflect some occasion-specific components due to shared occasion-specific perceptions (e.g., if many students in a class base their rating mostly on a few lessons in the past week instead of the specified referent period).

Research Questions

Although the empirical examination of these assumptions could have crucial implications for survey-based measurement of instructional quality, to our knowledge, no study has yet investigated the extent to which instructional variability influences the agreement between teacher and student ratings. Drawing inferences from the aforementioned research findings and theoretical assumptions, the present study's focus was twofold.

First, our study addressed the amount of variance that could be attributed to the consistent teaching practice component across several measurement occasions. On the one hand, teachers and students might become more familiar with the instrument across time (Koziol Jr & Burns, 1986; Lievens et al., 2007). When answering questions about features of regular instructional quality, this could help them answer the items more reliably. As a consequence,

the amount of the consistent teaching practice component at a single measurement occasion might increase. On the other hand, teachers' and students' commitment to the study might decrease due to the taking of repeated measurements and the implicated burden (Lanahan et al., 2005). This, in turn, might lead to an increase in residual variance in relation to variance that could be attributed to the consistent teaching practice component. Thus, for teacher and student ratings, both effects on the extent of the consistent teaching practice component (i.e., either an increase or a decrease) were believed to be possible. Our first research question thus examined the extent to which the consistent teaching practice component would actually be reflected by teacher and student ratings across several measurement occasions. Besides these "systematic" (perception-specific) components, also unsystematic effects may be relevant: In line with results from previous studies on observer ratings (Kane & Staiger, 2012; Praetorius et al., 2014) and student ratings (Rakoczy, 2008) we expected higher stability for classroom management compared to goal clarity and support of autonomy, because of its content-independence (which should lead to a low variability of the construct besides perceptual aspects) and the rather low degree of inference of the respective indicators. Comparing teacher and student ratings, we expected higher stability of the student ratings due to thethe aggregation of individual ratings at the class level, which should eliminate individual occasion-specific components (and measurement error).

Second, the present study investigated whether the association between teacher and student ratings would increase when aggregated measures of instructional quality were used instead of single-wave measurements. Based on prior research, we expected that teacher and student ratings would reflect stable as well as variable components of instructional quality (Curby et al., 2011; Kane & Staiger, 2012; Praetorius et al., 2014; Seidel & Prenzel, 2006). Because we assumed that the perception of variable components by teachers and students would be highly affected by perspective-specific referent periods (Clausen, 2002; Lanahan et al., 2005; Weinstein, 1985), we expected that the agreement between teacher and student ratings would increase when we aggregated ratings from several measurement occasions and separated the consistent from the variable component. Consequently, the association between the consistent components of the two perspectives was expected to be higher than the association between teacher and student ratings from a single measurement occasion. Based on empirical findings about construct-specific differences with regard to the association between teacher and student ratings (Clausen, 2002; Fauth et al., in press; Kunter & Baumert, 2006), however, we expected that these construct-specific differences would also remain when using aggregated measures of instructional quality: The largest association between

teacher and student ratings was expected for classroom management, because of the low content-dependency and a low degree of inference of the indicators, the latter leading to a low impact of self-serving strategies on teacher ratings.

Method

Procedure

The present analyses are part of a larger randomized experimental study about the effects of a teacher training program on students' self-regulated learning in math lessons[5] (see Werth et al., 2012). All teachers and half of the students in each class completed similarly worded items about various aspects of instruction. Data used in this study were based on teacher and student questionnaires that were administered three times across 13 weeks of schooling. The first measurement (t1) took place in February, the second measurement (t2) occurred 6 school weeks after the first measurement (i.e., the end of April), and the third measurement (t3) occurred in June (i.e., 6 or 7 school weeks after the second measurement occasion).

Sample

The present teacher sample consisted of 74[6] math teachers who taught fifth graders from the lowest school track ("Hauptschule") in Germany. The average age of the teachers was 45 years (*SD* = 11.74) and 69% were female. On average, they had 17 years of teaching experience (*SD* = 11.41) and about 78% taught the class not only in mathematics but were also the head teacher of the class. Due to time constraints, only about half of the students in each class completed questions about instructional quality, whereas the other students answered questions about their self-regulation. Thus, the student sample consisted of 686[7] students in these 74 classes. The average number of students who completed the instructional quality surveys was 9.27 per class. The average age was 11.76 years (*SD* = 0.74), and gender was distributed almost equally (51.7 % boys).

---

[5] The 4-day-long training delivered in this study was designed to develop teachers' ability to enhance students' self-regulation in math lessons. The study used an experimental design with random assignment of teachers to the treatment and control conditions. Because an effect of the treatment was expected only for the means of the applied scales but not for the teacher-student correlation, we treated the two groups as one. However, in order to control for differences in the initial status and potential unexpected treatment effects, we regressed all manifest means on a dummy-coded grouping variable (0 = control group, 1 = treatment group).

[6; 7] The actual sample size varied between measurement occasions and constructs from 70 – 74 teachers and 500 – 686 students. Except for the computation of Cronbach's alpha and the calculation of teacher and student descriptive statistics (i.e., sex, age, teaching experience), all analyses were conducted with Mplus 6.0 (Muthén & Muthén, 1998-2010) applying the full information maximum likelihood estimator (FIML; (Enders & Bandolos, 2001; Muthén & Muthén, 1998-2010) to deal with missing data. Thus, all analyses that were computed to address our research questions included information on 74 teachers and 686 students.

Measures

Both teacher and student questionnaires included a large number of scales, some of which had similar wording (cf. Werth et al., 2012). Mathematics teachers were asked to rate their own behavior in the mathematics class in question, whereas students were asked to rate their mathematics lessons/teacher. At t2 and t3, teachers and students were asked to keep in mind the previous 4 to 6 weeks, approximately corresponding to the time between measurement time points 1 and 2 and measurement time points 2 and 3. This design allows for the estimation of (mostly) *perception-specific variability* of teacher and student ratings (i.e., variability that may be seen as "error variance"). On the one hand, the referent periods are long enough to ensure quite comparable instructional quality in the different periods (Praetorius et al., for instance, found that nine lessons are enough to get a reliable measure even for highly variable constructs). On the other hand, the referent periods are short enough to (mostly) exclude "true" changes in instructional quality. Therefore, highly valid measures of such referent periods should be strongly correlated across measurement occasions.

For the purpose of the present investigation, we focused on a subset of three scales: *classroom management*, *goal clarity*, and *support of autonomy*, each being an exemplary scale of one the three domains of instructional quality. Using a group-oriented wording scheme, these scales were adapted to reflect the perspectives of the teachers and the students. Further information concerning the applied scales and items can be gathered from Table 1.

Classroom management. This scale taps the occurrence of disciplinary problems and belongs to the aforementioned domain of class organization. In the present investigation, classroom management was assessed with four items that were adapted from the PISA 2003 (Ramm et al., 2006) and COACTIV studies (Baumert et al., 2009) (sample item from the teachers' perspective: "There is hardly any disorder in this class."; corresponding item from the students' perspective: "With our math teacher, everything is calm and structured."). The reliability (Cronbach's alpha) of the classroom management scale ranged from $.86 \leq \alpha \leq .89$ for the teachers' perspective and from $.76 \leq \alpha \leq .80$ for the students' perspective. The relative agreement among students at the class level in terms of the intraclass-correlation was acceptable ($ICC_{t1} = .17$; $ICC_{t2} = .12$; $ICC_{t3} = .11$). **Goal clarity.** Goal clarity describes the extent to which the teacher clearly mentions his/her goals for the lesson, the week, or the year and can be regarded as an exemplary scale of the domain instructional support.

Table 1

Descriptives, Reliabilities, Intraclass Correlations, and Number of Items for the Applied Teacher and Student Scales

	t1			t2			t3			
Construct	*M(SD)*	Cronbach's α	*ICC*	*M(SD)*	Cronbach's α	*ICC*	*M(SD)*	Cronbach's α	*ICC*	Number of items
Teacher ratings										
Classroom management	2.37 (0.74)	.86	-	2.33 (0.74)	.89	-	2.38 (0.67)	.87	-	4
Goal clarity	3.38 (0.45)	.81	-	3.38 (0.46)	.84	-	3.24 (0.46)	.83	-	5
Support of autonomy	3.30 (0.40)	.61	-	3.31 (0.40)	.61	-	3.31 (0.40)	.72	-	4
Student ratings										
Classroom management	2.70 (0.29)	.76	0.17	2.74 (0.26)	.79	0.12	2.67 (0.25)	.80	0.11	4
Goal clarity	2.98 (0.20)	.75	0.09	2.93 (0.19)	.79	0.08	2.83 (0.19)	.79	0.08	5
Support of autonomy	3.02 (0.18)	.70	0.08	2.92 (0.15)	.77	0.05	2.82 (0.19)	.65	0.06	4

Note. Teacher sample: $N = 70 – 74$; Student sample: $N = 500 – 686$. Response format: 1 = *I don't agree at all*; 2 = *I don't agree*; 3 = *I agree*; 4 = *I completely agree.*

According to Seidel et al. (2005, p. 543), clear and transparent teaching goals should help students to identify with the learning contents and to integrate the teaching goals into their own learning goals. This, in turn, should enhance cognitive learning activities and learning motivation. In the present investigation, goal clarity was assessed with five items taken from the TRAIN study inventory (Jonkmann, Rose, & Trautwein, 2013) and the PYTHAGORAS study (Rakoczy, Buff, & Lipowsky, 2005) (sample item from the teachers' perspective: "I try to exactly explain the learning goals to my students."; corresponding items from the students' perspective: "Our teacher is able to explain his/her teaching goals well."). The reliability (Cronbach's alpha) of the scale ranged from $.81 \leq \alpha \leq .84$ for the teachers' perspective and from $.75 \leq \alpha \leq .79$ for the students' perspective. Almost 10% of the variability was at the class level ($ICC_{t1} = .09$; $ICC_{t2} = .08$; $ICC_{t3} = .08$).

Support of autonomy. This construct taps the extent to which students are allowed to choose their own strategies when solving problems and is one aspect of the domain emotional support. In line with the classification of support of autonomy as a high-inference construct whose rating is not "denotable [and] relatively objective" (Rosenshine, 1970, p. 281), the construct support of autonomy showed the lowest between-class variability ($ICC_{t1} = .08$; $ICC_{t2} = .05$; $ICC_{t3} = .06$), indicating a rather idiosyncratic perception of support of autonomy, which is also implied by the reference to the individual student for three of the four items of the scale from the student perspective. In the present investigation, support of autonomy was assessed with items (sample item from the teachers' perspective: "Before I help my students with solving a task, I encourage them to find the right solution by themselves."; sample item from the students' perspective: "Before my teacher helps me with solving a task, s/he encourages me to find the right solution on my own.") taken from the PALMA study inventory (Pekrun et al., 2002). Cronbach's α for the teachers' perspective ranged from $.61 \leq \alpha \leq .72$ and for the students' perspective from $.65 \leq \alpha \leq .77$.

Analyses

To answer our research questions, we first set up a manifest-latent state-trait model[8] for each construct based on the assumption that a teacher's performance with regard to classroom management, goal clarity, and support of autonomy should be partly consistent over time and partly variable (Curby et al., 2011; Kane & Staiger, 2012; Malmberg et al., 2010; Seidel & Prenzel, 2006) and that teacher and student ratings thus consist of a consistent teaching practice component and variable teaching practice component (the latter comprising both time-specific and residual variance). As depicted in Figure 1, all models were specified as two-level models (Hox, 2010; Lüdtke et al., 2008; Muthén & Muthén, 1998-2010) because our data were inherently organized in a hierarchical structure (i.e., students nested within classes). By using the Twolevel command in Mplus 6.0 (Muthén & Muthén, 1998-2010), the student covariance matrix was separated into within-class and between-class matrices, which here reflect components of idiosyncratic and shared student perceptions of instructional quality, respectively. To address missing data, we used the full information maximum likelihood estimator (FIML; Enders & Bandolos, 2001; Muthén & Muthén, 1998-2010). The postulated state-trait models were set up at the between-class level. Each of the models included two trait components representing the consistent component of teacher ratings (tcc) and student ratings (scc) of the construct over time. Manifest teacher mean scale scores (btt1, btt2, btt3) and manifest-latent student mean scale scores (bst1, bst2, bst3; i.e., indicators at the class level are represented by latent variables based on manifest mean scale scores labeled wst1, wst2, and wst3) of the three measurement occasions were used as indicators of the trait components. The residual variance components of the indicators can be interpreted as states (and measurement error) in the sense of a variable teaching component of the respective construct. According to the idea that aspects of instructional quality can vary to a certain extent and that some measurement occasions might be more representative of the consistent component of teaching than others, neither the factor loadings between the states and the respective trait nor the variances of the states were restricted. Based on these models, we calculated the percentage of variance in teacher and student ratings that could be attributed to

[8] Following the taxonomy of Marsh et al. (2009), the applied approach can be categorized as a manifest-latent partial correction model. This implies that the states of the estimated state-trait model are manifest in the sense that there is only one manifest score per construct instead of several items. Due to the lack of additional indicators per construct, measurement errors caused by the sampling of items could not be adjusted. However, at the same time, the class means at Level 2 were formed by latent aggregation of Level 1 constructs. As shown by Marsh et al. (2009), partial correction models are useful when samples sizes are small. Since doubly-latent models often suffer from convergence problems, the manifest-latent partial correction models are in some cases even more accurate than unbiased but unstable estimates gained by doubly-latent models (Marsh et al., 2009, p. 768).

the consistent teaching practice component and the measurement-occasion-specific component.[9]

To answer our second research question, relying on the aforementioned state-trait models, we estimated the correlations between teachers' and students' time-specific perceptions (see Figure 1; r_{t1}, $r_{t2,}$ r_{t3}). Furthermore, we estimated the correlation between the two factors from the teachers' and students' perspectives (see Figure 1; r_{aggr}). In order to compare these correlations with correlations drawn from single-wave data, we additionally estimated the correlation between the manifest-latent means for each measurement occasion ($r_{conventional\ t1}$, $r_{conventional\ t2,}$ $r_{conventional\ t3}$) in separate models.

Results

Extent of the Consistent Teaching Practice Component in Teacher and Student Ratings

Based on the state-trait-model (Figure 1), we first investigated the extent to which teacher and student ratings across several measurement occasions reflected consistent and the variable teaching practice components (first research question). For that purpose, we calculated the percentage of variance that could be attributed to the consistent teaching practice component and to a time-specific component (including measurement error) at each measurement occasion.

Concerning the underlying state-trait model, we determined that model fit was good for all models for each of the three constructs according to the conventional guidelines for acceptable fit (see Figure 1). The CFIs were greater than .98 for all models. Furthermore, the standardized factor loadings of the manifest means on the respective first-order factor ranged from $.73 < \lambda < .97$, indicating that most of the variance in the teacher and student ratings at the between-class level for different measurement occasions could be explained by either a consistent performance of the teacher or at least a stable perception of the respective instructional aspect by teachers and students.

[9] The percentage of total variance in teacher and student ratings that could be attributed to the consistent teaching practice component was obtained by first squaring the respective standardized factor loadings (see Figure 1; $\lambda_1 - \lambda_6$). For teacher ratings, the obtained percentages were then multiplied by the total variance and finally divided by an adjusted total variance. For student ratings, the obtained percentages were first multiplied by the between-class variance and then divided by an adjusted total variance. In order to calculate the percentage of total variance that could be attributed to the measurement-occasion-specific component, an almost identical procedure was applied. The adjustment of variance implied a correction of the amount of total variance for the amount of variance that could be attributed to the treatment condition. This procedure was aimed at the correction of a potential effect of the treatment.

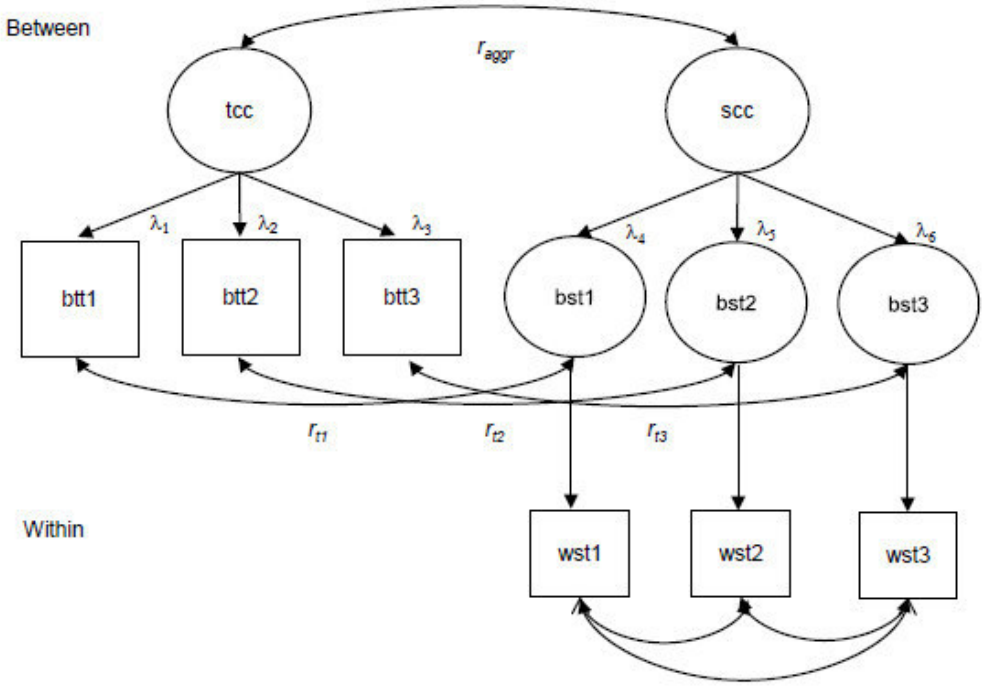

Figure 1. Two-level state-trait model correcting standard errors for clustering at the class level. Estimates are from Mplus 6.0 (Muthén & Muthén, 1998–2010), computed with the two-level command: within = individual student ratings; between = student ratings at the class level and teacher ratings. Fit indices for the three separate models: *Classroom management*: $\chi^2(6) = 13.35$, $p = .038$; CFI = .983; TLI = .932; RMSEA = .042; $SRMR_W = 0.003$; $SRMR_B = 0.049$. *Goal clarity*: $\chi^2(5) = 7.88$, $p = .163$; CFI = .994; TLI = .970; RMSEA = .029; $SRMR_W = 0.001$; $SRMR_B = 0.059$. *Support of autonomy*: $\chi^2(5) = 57.97$, $p = .981$; CFI = 1.000; TLI = 1.000; RMSEA = .000; $SRMR_W = 0.000$; $SRMR_B = 0.016$.

A closer look at the teacher ratings revealed that–in line with our hypothesis–the amount of variance in the consistent teaching practice component varied across time and constructs (see Table 2 and Figure 2). Whereas 69% (t_1) to 81% (t_3) of the variance in teachers' ratings of classroom management could be attributed to the consistent component, 53% (t_1) to 70% (t_2) of the variance in goal clarity and 61% (t_1) to 71% (t_3) of the variance in support of autonomy could be attributed to the consistent component. In accordance with our hypothesis the largest average amount of variance of the consistent component was found for the construct classroom management. It should be noted, that this result may be a consequence of the low degree of inference of the items, which may be the reason for the higher internal consistency of the classroom management scale (see Cronbach's α in Table 1).

Table 2

Percentages of Total Variance in Teacher and Student Ratings that could be Attributed to the Consistent Teaching Practice and the Variable Teaching Practice Component (all Measurement Occasions; t1, t2, t3)

	Classroom management			*Goal clarity*			*Support of autonomy*		
Teacher ratings	t1	t2	t3	t1	t2	t3	t1	t2	t3
Consistent component	69	79	81	53	70	62	61	64	71
Variable component	31	21	19	47	30	38	39	36	29
Student ratings	t1	t2	t3	t1	t2	t3	t1	t2	t3
Within-class variance	84	89	89	92	92	94	93	95	94
Consistent component	15	6	8	8	7	6	7	4	6
Variable component	1	5	2	0	0	0	0	0	0

Note. Teacher sample: $N = 74$; Student sample: $N = 686$. Percentages may not total 100 due to rounding

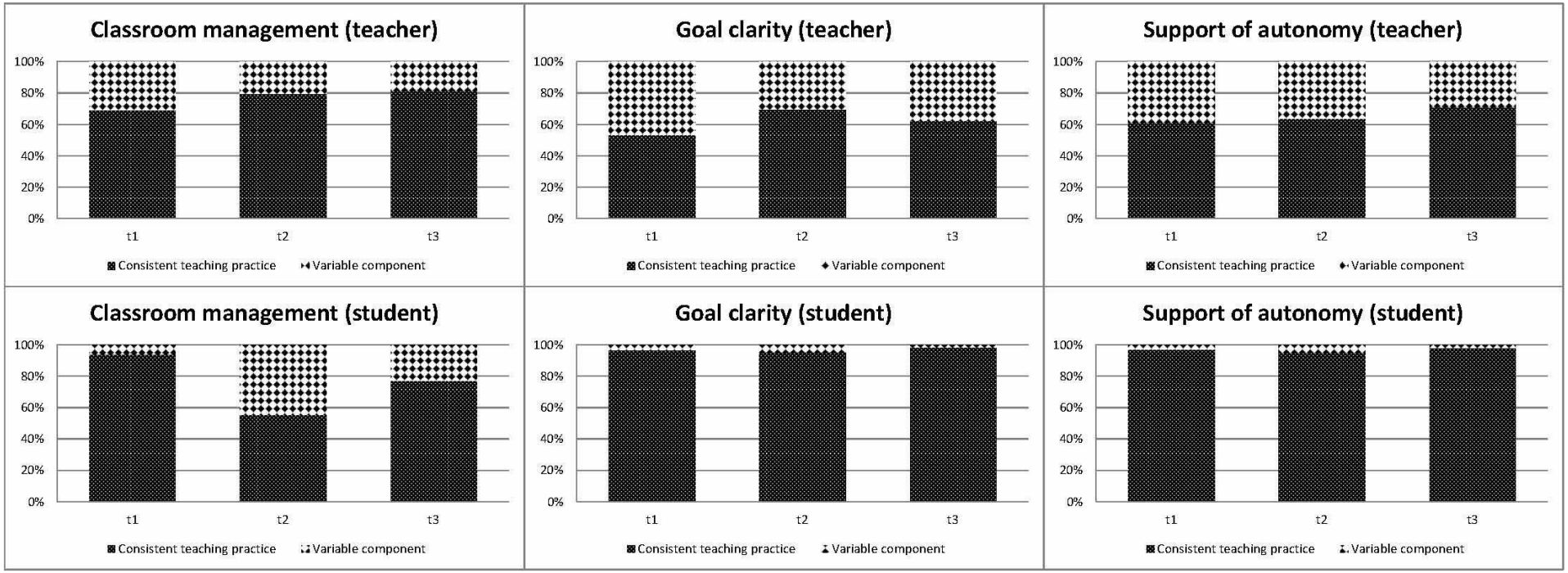

Figure 2. Percentages of variance on class level that can be attributed to the consistent teaching practice component of teacher and student ratings at all measurement occasions (t1, t2, t3). We also controlled for potential effects of the treatment condition. Because on average only 1% of the variance in teacher ratings and 5% of the between-class variance in student ratings could be attributed to the treatment, this component is not included in Figure 2.

With regard to student ratings of the three constructs the following results were found: 94% (t_1), 55% (t_2), and 77% (t_3) of the between-class variance of classroom management was covered by the consistent teaching practice component. By contrast, as depicted in Figure 2, the amount of between-class variance in goal clarity and support of autonomy that could be attributed to the consistent component ranged from 95% to 99% (goal clarity: t1: 97%; t2: 95%; t_3: 99%; support of autonomy: t_1: 97%; t_2: 95%; t_3: 98%). In line with our hypothesis we can thus state that also in case of the student ratings the amount of consistent variance varied across measurement time points. The findings revealed, as expected, that classroom management was the construct with the highest stability in relation to the total variance of the construct (t_1: 15%, t_2: 6%, t_3: 8%; see Table 2 which depicts within- and between-class variance components. However, contrast to our hypothesis, the amount of variance at the between-class level of the consistent teaching practice component was largest for the content-dependent constructs (goal clarity, support of autonomy). Comparing teacher and student ratings it can be seen that in line with our hypothesis at the between-class level the consistent teaching practice component accounted for a much larger amount of variance (see Figure 2 which only shows between-class variance).

In summary, we found that the consistent components in both teacher and student ratings for all constructs explained a large amount of variance, which, for aggregated student ratings, partly approximated 100 %. Construct-specific differences regarding the consistent component seem rather to be a consequence of the degree of inference of the indicators than of the content-dependency of the construct.

Correlations between Nonaggregated and Aggregated Teacher and Student Ratings

Our second research question was aimed at investigating the extent to which the correlation between teacher and student ratings would increase when correlations between aggregated measures of instructional quality were analyzed instead of correlations between single time points. We first estimated correlations between manifest teacher and student ratings at measurement time points 1, 2, and 3. Second, based on state-trait models (as depicted in Figure 1), we estimated the correlation between the consistent components of teaching from the two perspectives and the correlations between the residuals from the two perspectives that corresponded to the cross-sectional correlations that were estimated before.

With regard to correlations for cross-sectional teacher and student data, the agreement between teacher and student ratings varied across constructs and time (see Table 3).

Table 3

Agreement between Nonaggregated and Aggregated Teacher and Student Ratings of Instructional Quality at all Measurement Occasions (t1, t2, t3)

	Conventional correlation at the class level	Correlation in state-trait model (Figure 1)
Classroom management	$r_{conventional\ t1}$ = .72**	$r_{t1} = ns$
	$r_{conventional\ t2}$ = .58**	$r_{t2} = ns$
	$r_{conventional\ t3}$ = .77**	$r_{t3} = ns$
	-	$r_{aggr} = 1.00$
Goal clarity	$r_{conventional\ t1} = .45^{**}$	$r_{t1} = ns$
	$r_{conventional\ t2} = .35^{\dagger}$	$r_{t2} = ns$
	$r_{conventional\ t3} = ns$	$r_{t3} = ns$
	-	r_{aggr} = 0.50**
Support of autonomy	$r_{conventional\ t1} = ns$	$r_{t1} = ns$
	$r_{conventional\ t2} = ns$	$r_{t2} = ns$
	$r_{conventional\ t3} = ns$	$r_{t3} = ns$
	-	$r_{aggr} = ns$

Notes. Teacher sample: N = 74; Student sample: N = 686.
[a] The estimate was out of range (r = 1.031; p = .000) and thus restricted to r = 1.00.[6]
† $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Whereas the highest correlations, as expected, were found for classroom management, $r_{conventional\ t1} = .72$, $r_{conventional\ t2} = .58$, $r_{conventional\ t3} = .77$, a lower correlation was found for goal clarity at the first measurement occasion ($r_{conventional\ t1} = .45$) and no statistically significant correlations were found for the second and third time points. For support of autonomy, there was no statistically significant association between teacher and student ratings at any measurement occasion.

Considering the agreement between teachers and students in the state-trait model, the construct-specific pattern of correlations also persisted when aggregating teacher and student ratings across time. As expected the the largest correlation was found for classroom management. Whereas the trait components of classroom management showed a perfect correlation[10], the correlations between the variable states (i.e., the residuals) were not statistically significant in the state-trait model (see Figure 1). Also, the agreement between teachers' and students' ratings of goal clarity increased to $r_{aggr} = .50$, whereas the correlations between the variable components (residuals) were no longer statistically significant. Concerning the construct support of autonomy, neither the consistent components nor the variable components (residuals) showed statistically significant correlations.

Discussion

The current study addressed the questions of whether measures of instructional quality are able to capture consistent and variable teaching practice components of instructional quality across several measurement occasions and the extent to which the aggregation of teacher and student ratings of instructional quality from several measurement occasions would lead to higher convergent validity of these measures. The results showed that at the class level the largest variance component represents consistent teaching practice for all constructs in the present study rated by teachers as well as students. The largest relative agreement between both perspectives for a single measurement occasion was found for classroom management, while for support of autonomy no statistically significant correlations were found. The same pattern was found regarding the relative agreement for the consistent components at the class

[10] The standardized estimate in the original model was out of range ($r = 1.031$, $p = .000$), but as the 95% confidence interval covered the maximum possible value of $r = 1.00$, the respective model parameter was restricted to this value. Besides small sample sizes (Anderson & Gerbing, 1984; Hox, 2010), model misspecifications were shown to be the potential causes of improper solutions (Chen, Bollen, Paxton, Curran, & Kirby, 2001). Although the model fit here was acceptable, we checked the modification indices for possible post hoc modifications. The largest chi-square drop was found for the residual correlation between the Wave 1 student scale and the Wave 3 teacher scale. A model with this additional parameter resulted in an excellent model fit, $\chi^2(4) = 6.304$, $p = .178$, CFI = .995, TLI = .968, RMSEA = .029, $SRMR_W = 0.002$, $SRMR_B = 0.038$, and the correlation between the two trait factors was estimated as $r = .96$ ($p < .001$).

level (aggregated over three measurement occasions): A perfect correlation was found for classroom management, a medium correlation for goal clarity, and no statistically significant correlations were found for support of autonomy.

Consistent Teaching Practice in Teacher and Student Ratings

With regard to our first research question findings from our study indicate that teacher and student ratings across several measurement occasions contain a consistent teaching practice component as well as a component reflecting time-specific and residual variance components—just as was found for the observer ratings in the MET project (Kane & Staiger, 2012). Concerning the reliability of student ratings, our findings showed distinctive differences between the within-class and between-class levels. For the between-class level, student ratings were highly representative of the consistent teaching practice component (55% to 99% of between-class variance). With regard to the percentage of total variance that could be attributed to the consistent component, however, single-wave student ratings were less representative of the teacher's general practice in the classroom than observer ratings (student ratings [single measurement occasion]: 4% to 15%; observer ratings [single measurement occasion]: 23% to 32%). Concerning the assumed influence of repeated measurements on teacher and student ratings, we can draw two conclusions. First, teachers' ratings seem to be rather unaffected by repeated measurements i.e., although we could recognize a slight increase in the consistent teaching practice component from t1 to at t3 in the case of classroom management and support of autonomy this was not the case for the scale goal setting which showed the highest value for t2. In summary we can state that there was either no systematic change from t1 to t3 or even a slight increase in the consistent teaching practice component. Second, the variance in students' ratings that can be attributed to the consistent teaching practice component was larger than the corresponding component of teacher ratings and varied across time, constructs, and levels. Whereas the amount of variance at the class level that could be attributed to the consistent component remained rather stable, the percentage of total variance that could be explained by the consistent component decreased over time. In conclusion, support was only found for an influence of repeated measurements on the amount of within-class variance but not on the consistent and variable teaching practice components. In line with research by den Brok, Fisher, Rickards, and Bull (2006), this increase of within-class variance might indicate that repeated measurements do not necessarily affect all students to an equal extent but have differential effects on individual students. They showed that not only might the shared perception of instructional quality be an

important predictor of students' learning but that individual differences in perceptions of instructional quality, which are reflected by the amount of within-class variance, might also predict learning. Based on our results, it seems reasonable that students actually became accustomed to the questionnaires and that their answers became more representative of the individually perceived support of autonomy, which, in turn, led to an increase in within-class variance in student ratings. With regard to the differences between the constructs the study indicates that the degree of inference, the content-dependency of the items and the choice of the addressee (the individual student vs. the class) are reflected in the differences between the constructs.

Relative Agreement between Teacher and Student Ratings aggregated over Measurement Occasions

With regard to our second research question that addressed the correlation between aggregated teacher and student ratings of instructional quality, our study showed that teacher-student agreement concerning classroom management increased to a perfect correlation. Furthermore, the correlation between teacher and student ratings of goal clarity increased to some extent. Contrary to our hypothesis, the agreement between teacher and student ratings of support of autonomy did not increase when relying on aggregated measures. The assumption that construct-specific differences between the correlation of teacher and student ratings remain after aggregation was confirmed: The highest correlation was still found for classroom management and the lowest correlation was found for support of autonomy. As students' perceptions of support of autonomy were almost uncorrelated even within classrooms, however, the explanation that the gap between teacher and student ratings is mostly a consequence of instructional variability in combination with different referent periods—which are both influences that occur at the class level—seems to be insufficient. Our finding that the correlations between teacher and student ratings of classroom management and goal clarity increased when aggregating across several measurement occasions might indicate that the influence of instructional variability and referent periods on teacher-student agreement only comes into play in case that an agreement between teacher and student perspectives can be expected at all.

Concerning the validity of the extracted components it might be recognized that the increased correlations between the aggregated teacher and student ratings indicate an increased convergent validity at least in case of the consistent teaching practice component. Unfortunately, in case of the variable teaching component this indicator of validity is missing

because the correlations between the variable components in teacher and student ratings were not statistically significant. Yet, despite this lack of external validation and in line with the common practice to ask teachers and students for an averaged impression and assuming teaching practice to be quite stable our study treats the variable component as a measurement error that reflects occasion- and rater-specific bias like e.g. primacy-recency-effects and thus something that was not intended to measure when assessing consistent teaching practice. Consequently, this variable component was not further interpreted and the analyses were focused on the association between the consistent components of teachers' and students' ratings. However, with regard to findings about the variability of instructional quality (Curby et al., 2011; Kane & Staiger, 2012; Praetorius et al., 2014; Seidel & Prenzel, 2006), future studies should also address the importance of consistent and variable aspects of teaching for students' learning.

Limitations and Further Research

Although several research projects such as the MET project (Kane & Staiger, 2012) have recently tried to enhance the quality of measures of instructional quality, no study has yet compared teacher and student ratings of instructional quality in a longitudinal design. Based on the assumption that consistent and variable components of teacher and student ratings are indicators for instructional variability the study could show that instructional variability is also reflected in teacher and student ratings. By showing that the aggregation of the ratings from the two perspectives helps to alleviate this influence, our study is able to make an important contribution to research about instructional quality in mathematics education.

Yet, the present study also raises some theoretical and methodological issues. For instance, the selected scales differ greatly concerning the degree of inference asking on the one hand for observations of teacher practices (e.g. whether a teacher clearly explains what students should learn) and on the other hand for subjective impressions (e.g. whether the teacher allows the students to apply own strategies). As the degree of inference may be assumed to have a large impact on the (relative) agreement between student and teacher ratings (Kunter & Baumert, 2006; Fauth et al., in press), correlations for scales within the same domain may vary due to different degrees of inference of the respective indicators. Therefore, the restriction to one scale for each domain might probably have implications for the comparability of our results to studies that assessed the three domains with several scales for each domain. For instance, our finding that emotional support was the least stable domain is contradictive to the finding of Curby et al. (2011) who found this domain to be the most

stable. Whereas Curby et al. (2011) refer to several scales within the domain emotional support that differ with regard to their stability, some being more stable and few being less stable than the single constructs from the other domains our results are only based on the scale support of autonomy. Whether different, rather low-inferent scales from the domain emotional support would have led to different results would be worth investigating in further research.

Another shortcoming of the study is that, unfortunately, the separation of time-specific variance from residual variance was not possible due to the decreasing intraclass correlation and the restricted number of classrooms in our study that made it impossible to analyze data at the item level (i.e., to set up doubly-latent instead of manifest-latent state-trait models; (Marsh et al., 2009)). Because of this limitation, the reliability of teacher ratings might be reduced due to sampling errors, which, in turn, might lead to an underestimation of correlations between aggregated teacher and student ratings (Marsh et al., 2009). With regard to future research, we thus suggest that the findings that we presented in this study be demonstrated with other data sets that also allow for detailed analyses at the item level.

Finally, we have to mention that all the reported impressions of a de- or increase of percentages of variance only have face-validity and are not confirmed by statistical tests. The reason for this is the small sample size at the class level (combined with rather low ICCs for the student perceptions) which makes it impossible to reasonably estimate such parameters in large models with several constructs and measurement occasions simultaneously.

Overall, our findings point to the value of finding ways to improve the measures that we apply when estimating the success of instructional reforms or when estimating the influence of instructional quality on students' achievement gains. Although the present study will not help to fully overcome the depicted validity problem, our findings indicate that teacher and student ratings are to some extent influenced by instructional variability and that the gap between the two perspectives can be reduced by aggregating ratings across several measurement occasions. The study thus helps to explain the low agreement between teacher and student ratings in cross-sectional studies and might reduce this gap in future studies that provide more than two measurement occasions.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science 333*(6045), 1034-1037.

Anderson, J., & Gerbing, D. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173. doi: 10.1007/bf02294170

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U. et al. (2009). Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente. [Professional competence of teachers, cognitively activating instruction, and the development of students' mathematical literacy (COACTIV): Documentation of the instruments]. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Blum, W., Brunner, M., Krauss, S., Kunter, M., & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. [Mathematics teaching from the perspective of the PISA students and their teachers]. In M. Prenzel, J. Baumert, W. Blum, R. Lerhmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (pp. 314-354). Münster: Waxmann-Verlag.

Caroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies. *Sociological Methods & Research, 29*(4), 468-508.

Clausen, M. (2002). Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt - und Kriteriumsvalidität. [Quality of instruction: A matter of perspective?]. Münster, Westfalen u.a.: Waxmann.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*(1), 113-143. doi: 10.3102/003465430298563

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J. et al. (2011). Within-day variability in the quality of classroom interactions during third and

fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal, 112*(1), 16-37.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51-85.

den Brok, P., Fisher, D., Rickards, T., & Bull, E. (2006). Californian science students' perceptions of their classroom learning environments. *Educational Research and Evaluation, 12*(1), 3-25. doi: 10.1080/13803610500392053

Desimone, L., Smith, T., & Frisvold, D. (2009). Survey measures of classroom Instruction: Comparing student and teacher reports. *Educational Policy*. doi: 10.1177/0895904808330173

Enders, C. K., & Bandolos, D. L. (2001). The relative performance to full information maximum likelihood estimation for missing data in structural equation models. . *Structural Equation Modeling, 8*(3), 430–457.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., Büttner, G. (in press). Grundschulunterricht aus Schüler-, Lehrer und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. [Instructional quality in elementary school from the perspective of students, teachers, and observers: Associations and prediction of learning outcomes.]. *Zeitschrift für Pädagogische Psychologie*.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K. et al. (2010). Middle school mathematics professional development impact study: Findings after the first year of implementation. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hamre, B. K., Pianta, R. C., & Chomat-Mooney, L. (2009). Conducting classroom observations in school-based research. In L. M. Dinella (Ed.), *Conducting science-based psychology research in schools.* (pp. 79-105). Washington, DC US: American Psychological Association.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. NY: Routledge.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2013, August 12). Classroom Composition, Classroom Management, and the Relationship Between Student Attributes and Grades. [Advance online publication]. *Journal of Educational Psychology*, 106(1). doi: 10.1037/a0033829

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications.* (2nd ed.). New York: Routledge.

Jonkmann, K., Rose, N., & Trautwein, U. (Eds.) (2013). Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen. Abschlussbericht für die Länder Baden-Württemberg und Sachsen. Tübingen: Universität Tübingen.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. In B. M. G. Foundation (Ed.), *MET Project Research Paper*.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". [Quality dimensions and effectiveness of mathematics instruction. Theoretical background and selected findings of the Pythagoras project]. In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule.* (pp. 127-146). Münster, Westfalen u.a.: Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom.* (pp. 137-160). Münster u.a.: Waxmann.

Koziol Jr, S. M., & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research, 79*(4).

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231-251. doi: 10.1007/s10984-006-9015-7

Lanahan, L., McGrath, D. J., McLaughlin, M., Burian-Fitzgerald, M. A., & Salganik, L. (2005). Fundamental problems in the measurement of instructional processes: Estimating reasonable effect sizes and conceptualizing what is important to measure. Washington, DC: American Institutes for Research.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*(6), 1672-1682. doi: 10.1037/0021-9010.92.6.1672

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527-537. doi: 10.1016/j.learninstruc.2008.11.001

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203-229. doi: 10.1037/a0012869

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*(4), 916-932. doi: 10.1037/a0020920

Marsh, H. W., & Grayson, D. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling: A Multidisciplinary Journal, 1*(2), 116-145. doi: 10.1080/10705519409539968

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. et al. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764-802. doi: 10.1080/00273170903333665

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, Mass.: Center for Study of Testing, Evaluation, and Educational Policy. Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1998). Mathematics and science achievement in the final year of secondary school: IEA's Third International Mathematics and Science Study (TIMSS). Chestnut Hill, Mass.: Center for Study of Testing, Evaluation, and Educational Policy. Boston College.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition.*: Los Angeles, CA: Muthén & Muthén.

Organisation for Economic Co-operation and Development [OECD]. (1999). *Measuring Student Knowledge and Skills. A New Framework for Assessment*. Paris: OECD.

Pekrun, R., Götz, T., Jullien, S., Zirngibl, A., Hofe, R. v., & Blum, W. (2002). *Skalenhandbuch PALMA: 1. Messzeitpunkt (5. Klassenstufe) [Scale handbook PALMA, First Point of Measurement (Grade 5)].* Universität München: Institut Pädagogische Psychologie.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (Eds.). (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore: Paul H. Brookes.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2-12.

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). Befragungsinstrumente. [Questionnaires]. In E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 1).* Frankfurt a.M.: GFPF/DIPF.

Rakoczy, K. (2008). Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern [Motivational support in mathematics lessons: Instruction from the perspectives of learners and observers]. Münster, Germany: Waxmann.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D. et al. (Eds.). (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente. [Questionnaires]*. Münster: Waxmann.

Rosenshine, B. (1970). Evaluation of Classroom Instruction. *Review of Educational Research, 40*(2), 279-300. doi: 10.3102/00346543040002279

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

Seidel, T., & Prenzel, M. (2006). Stability of teaching patterns in physics instruction: Findings from a video study. *Learning and Instruction, 16*(3), 228-240. doi: 10.1016/j.learninstruc.2006.03.002

Seidel, T., Rimmele, R., & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction 15*(6), 539-556. doi: 10.1016/j.learninstruc.2005.08.004

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). doi: 10.3102/0034654307310317

Tsai, Y.-M., Kunter, M., Lüdtke, O., Trautwein, U., & Ryan, R. M. (2008). What makes lessons interesting? The role of situational and individual factors in three school subjects. *Journal of Educational Psychology*, 100(2), 460-472. doi: 10.1037/0022-0663.100.2.460

Urdan, T. (2004). Using multiple methods to assess student's perceptions of classroom goal structures. *European Educational Psychologist 9*, (4), 222–231. doi: 10.1027/1016-9040.9.4.222

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*(3), 249-294. doi: 10.3102/00346543063003249

Weinert, F. E., Schrader, F. W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Research, 13*, 895-914.

Weinstein, R. S. (1985). Student mediation of classroom expectancy effects. In J. B. Dusek (Ed.), *Teacher expectancies.* Hillsdale, NJ: Erlbaum.

Werth, S., Wagner, W., Ogrin, S., Trautwein, U., Friedrich, A., Keller, S. et al. (2012). Förderung des selbstregulierten Lernens durch die Lehrkräftefortbildung «Lernen mit Plan»: Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität. [Teaching Teachers how to teach Self-Regulated Learning: Effects of a Training Program on the Promotion of Self-Regulation and Instructional Quality]. *Zeitschrift für Pädagogische Psychologie, 26*(4), 291-305. doi: 10.1024/1010-0652/a000080

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education*, 8(1), 47–58

Werth, S.; Wagner, W.; Ogrin, S; Trautwein, U.; Friedrich, A.; Keller, S. et al. (2012): Förderung des selbstregulierten Lernens durch die Lehrkräftefortbildung "Lernen mit Plan": Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität. Zeitschrift für Pädagogische Psychologie, 26 (4), 291-305.[a]

DOI 10.1024/1010-0652/a000080

Z. Pädagog. Psychol. 26 (4) © 2012 Verlag Hans Huber, Hogrefe AG, Bern

3

Förderung des selbstregulierten Lernens durch die Lehrkräftefortbildung „Lernen mit Plan“: Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität

Werth, S., Wagner, W., Ogrin, S., Trautwein, U., Friedrich, A., Keller, S. et al. (2012). Förderung des selbstregulierten Lernens durch die Lehrerfortbildung „Lernen mit Plan“: Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität. *Zeitschrift für Päagogische Psychologie, 26*, 291–305.

Zusammenfassung

Lehrkräftefortbildungen gelten als nachhaltige Maßnahme der Selbstregulationsförderung bei Schülerinnen und Schülern. Noch immer besteht jedoch Forschungsbedarf zur Implementation der Selbstregulationsförderung im Unterricht und deren Auswirkung auf die allgemeine Unterrichtsqualität. Die viertägige Fortbildung „Lernen mit Plan“ vermittelt Haupt- und Werkrealschullehrkräften Strategien zur Selbstregulationsförderung im Mathematikunterricht. Im vorliegenden Beitrag wird untersucht, inwieweit „Lernen mit Plan“ die fokal trainierte Förderung der Selbstregulation (selbstregulationsspezifische Unterrichtsqualität) sowie die allgemeine Unterrichtsqualität beeinflusst. Insgesamt 75 Lehrkräfte wurden randomisiert einer Experimentalgruppe (47 Lehrkräfte) sowie einer Wartekontrollgruppe (28 Lehrkräfte) zugewiesen. Die selbstregulationsspezifische sowie allgemeine Unterrichtsqualität wurden in einem prä-, post- und follow-up-Design über Lehrer- und Schülerfragebögen erhoben. „Lernen mit Plan“ ging aus Lehrer- und Schülerperspektive mit einer verstärkten Selbstregulationsförderung und aus Schülerperspektive zusätzlich mit positiven Effekten auf die allgemeine Unterrichtsqualität einher.

Schlüsselwörter: Selbstregulation, Lehrkräftefortbildung, Unterrichtsqualität

Einleitung

Selbstreguliertes Lernen ist ein proaktiver Prozess, der Elemente wie Zielsetzung, Planung, den selektiven Einsatz von Lernstrategien und das Überwachen des eigenen Lernprozesses durch den Lernenden umfasst (Zimmerman, 2008). Dem selbstregulierten Lernen wird eine hohe Bedeutung beigemessen, da selbstreguliertes Lernen sowohl ein wichtiges, eigenständiges Ziel von schulischem Unterricht darstellt als auch als eine Ressource ist, die hilft, weitere Unterrichtsziele zu erreichen (Dignath & Büttner, 2008; Hattie, Bigs & Purdie, 1996; Glaser, Keßler & Brunstein, 2009; Perels, Gürtler & Schmitz, 2005; Trautwein & Köller, 2003; Vohs & Baumeister, 2011; Zimmerman, 2001).

Lehrkräftefortbildungen als Maßnahme der Lehrerprofessionalisierung kommt eine wichtige Rolle dabei zu, der zunehmenden Bedeutung der Förderung der Selbstregulation im Schulunterricht gerecht zu werden (Dignath & Büttner, 2008; Landmann & Schmitz, 2007; vgl. auch Borko & Putnam, 1995). Bisher vorliegende Studien zur Förderung der Selbstregulation deuten darauf hin, dass durch Lehrkräftefortbildungen die Selbstregulationskompetenzen von Schülerinnen und Schülern unterschiedlicher Altersgruppen und Schulformen positiv beeinflusst werden können (Dignath & Büttner, 2008; Hattie et al., 1996; Landmann & Schmitz, 2007).

In zahlreichen Studien bleibt jedoch offen, ob die entsprechende Fortbildung überhaupt dazu führte, dass tatsächlich eine Implementation der Selbstregulationsförderung im Unterricht stattfand (vgl. Fuchs & Fuchs, 2001; Hager & Hasselhorn, 2000; Marks et al., 1993). Zudem wurde bislang die konzeptuell und praktisch relevante Frage ausgeblendet, inwieweit erfolgreiche Fortbildungsmaßnahmen neben der selbstregulationsspezifischen Unterrichtsqualität auch die allgemeine Unterrichtsgestaltung von Lehrkräften und damit einen wesentlichen Bestandteil des Wirkungsgefüges lehrerzentrierter Interventionsmaßnahmen beeinflussen (Borko & Putnam, 1995).

Der vorliegende Beitrag basiert auf der Interventionsstudie „Lernen mit Plan“ zur Förderung von Selbstregulationsfähigkeiten im Mathematikunterricht der fünften Jahrgangsstufe von Haupt- und Werkrealschulen. In diesem Interventionsprojekt wurden Lehrkräfte darin trainiert, ihren Schülerinnen und Schülern die systematische Anwendung von Selbstregulationsstrategien zu vermitteln. Unter Verwendung der Angaben aus den Lehrer- und Schülerfragebögen untersuchen wir, ob sich „Lernen mit Plan“ in gewünschter Weise auf die Umsetzung der Trainingsinhalte zur Förderung von Selbstregulationsstrategien im Unterricht ausgewirkt hat. Darüber hinaus prüfen wir, ob sich die Fördereffekte des Trainings auf die fokalen Trainingsinhalte beschränken oder ob sich auch Veränderungen hinsichtlich

der allgemeinen Unterrichtsqualität (Qualität von Klassenführung, kognitiver Aktivierung und Schülerorientierung; vgl. Klieme, Schümer & Knoll, 2001) finden lassen.

Förderung der Selbstregulation von Schülerinnen und Schülern durch Lehrkräftefortbildungen

Zahlreiche Studien weisen darauf hin, dass das selbstregulierte Lernen erfolgreich gelehrt und gelernt werden kann (vgl. Dignath & Büttner, 2008). Die Vermittlung von Selbstregulation führt dabei besonders in Kombination mit Fachinhalten zu einer Verbesserung der Selbstregulation, aber auch der fachlichen Leistung (Glaser et al., 2009; Hattie et al., 1996; Labuhn, Zimmerman & Hasselhorn, 2010).

Bei der Förderung der Selbstregulation wird oftmals zwischen *expliziter* und *impliziter* Förderung unterschieden (vgl. Dignath, 2009). Unter expliziter Förderung wird dabei die bewusste Besprechung und Einübung der im zugrundliegenden Selbstregulationsmodell benannten kognitiven, metakognitiven und motivationalen Strategien mit den Schülerinnen und Schülern verstanden. Unter impliziter Förderung hingegen wird die Gestaltung einer die Selbstregulation der Schülerinnen und Schüler unterstützenden Lernumgebung (vgl. Dignath, 2009) gefasst, worunter nicht nur die Lerninhalte, sondern auch selbstregulationsaktivierende Hinweise durch die Lehrkraft, die behandelten Aufgaben und die angewandten Unterrichtsmethoden verstanden werden. Eine Kombination beider Fördermaßnahmen scheint hinsichtlich einer nachhaltigen Strategieförderung und Leistungsverbesserung seitens der Schülerinnen und Schüler besonders erfolgversprechend zu sein (vgl. Kistner et al., 2010).

Wenngleich Studien zur Förderung der Selbstregulation darauf hindeuten, dass durch Lehrkräftefortbildungen die Selbstregulationskompetenzen von Schülerinnen und Schülern unterschiedlicher Altersgruppen und Schulformen positiv beeinflusst werden können, sind die erzielten Effekte jedoch meist geringer als die Effekte von Schülertrainings durch externe Trainer (Dignath & Büttner, 2008). Zudem variieren die Befunde teilweise stark über verschiedene Studien hinweg (Dignath & Büttner, 2008). Die Ursachen für das Ausbleiben von Effekten und divergente Ergebnismuster könnten vielfältiger Natur sein. Zum einen könnten die nur schwach ausgeprägten Interventionseffekte auf Defizite der entsprechenden selbstregulationsbezogenen Lehrkräftefortbildungen hinweisen (Dignath & Büttner, 2008). So ist beispielsweise in vielen Studien unklar, inwieweit die Lehrkräfte die Trainingsinhalte im Anschluss an eine Fortbildung tatsächlich im Unterricht umgesetzt haben (Fuchs & Fuchs, 2001; Marks et al., 1993). Verschiedene Metaanalysen und Überblicksartikel zur Wirksamkeit von Lehrkräftefortbildungen haben hierbei darauf hingewiesen, dass zu kurze

Lehrkräftefortbildungen mit einem Ausbleiben der antizipierten Effekte assoziiert sind (Darling-Hammond, Wei, Andree, Richardson & Orphanos, 2009; Lipowsky, 2010). Zum anderen darf angenommen werden, dass das Fehlen von vorkonzipierten Lernmaterialien für Schülerinnen und Schüler den Lehrkräften einen effektiven Umstieg auf ein verändertes Lehrkonzept erschwert (vgl. De Corte, 2000; Lipowsky, 2010).

Insgesamt erweist sich der Transfer von Fortbildungsinhalten in die Unterrichtspraxis als ein komplexer und anspruchsvoller Prozess (Lipowsky, 2010), der einen bedeutsamen Eingriff in die Unterrichtsgestaltung darstellt. Um die Umsetzung der Trainingsinhalte von Lehrkräftefortbildungen im Unterricht zu fördern, stellen daher eine Reihe von Fortbildungen Unterrichtsmaterialien und geskriptete Unterrichtsentwürfe bereit (vgl. De Corte, 2000; Landmann & Schmitz, 2007; Mokhlesgerami, Souvignier, Rühl & Gold, 2007; Perels, Dignath & Schmitz, 2009). Die entsprechenden Trainings – so auch das in dieser Studie vorgestellte Training „Lernen mit Plan“ – sind damit zumindest teilweise manualbasiert. Es stellt sich entsprechend die Frage, welche Konsequenzen dies für die Umsetzung der fokalen Trainingsinhalte (Souvignier & Trenk–Hinterberger, 2007), aber auch für die allgemeine Unterrichtsqualität hat (Borko & Putnam, 1995). Im Hinblick auf die Umsetzung der fokalen Trainingsinhalte kann man aufgrund der Befunde bisheriger Lehrkräftefortbildungen zur Förderung der Selbstregulation davon ausgehen, dass eine relativ lange Trainingszeit, ein hoher Strukturierungsgrad der Fortbildungsinhalte sowie die Bereitstellung von Unterrichtsmaterialien zu einer stärkeren Umsetzung der Trainingsinhalte im Unterricht führt (vgl. Dignath & Büttner, 2008). Weniger klar vorhersagbar sind die Effekte auf nicht direkt trainierte Unterrichtscharakteristika (vgl. Borko & Putnam, 1995). Einerseits könnte sich die Bereitstellung vorkonzipierter Unterrichtsstunden, vorbereiteter Unterrichtsmaterialien und Empfehlungen zur selbstregulationsfördernden Bearbeitung von Aufgaben im Unterricht positiv auf die allgemeine Unterrichtsgestaltung von Lehrkräften auswirken. So wäre es beispielweise denkbar, dass Lehrkräfte, in deren Unterricht aufgrund mangelnder Strukturierung häufig Zeit nicht effizient genutzt wird, durch die Stundenentwürfe Anregungen erhalten, wie Sie Ihren Unterricht zeitlich effizienter gestalten können. Des Weiteren ist es möglich, dass die individualisierenden Elemente der Schülerförderung zu positiven Transfereffekten auf weitere Aspekte der Schülerorientierung führen. Andererseits wäre es jedoch auch möglich, dass die Konzentration der Lehrkraft auf die Selbstregulationsförderung zu einer weniger leistungs- oder motivationsförderlichen Veränderung der allgemeinen Unterrichtsgestaltung führt. So gehen Ophardt und Thiel (2008) in Anlehnung an Gruehn (1995) davon aus, dass eine multiple Zielerreichung (z.B. die

gleichzeitige Förderung der Leistung und der Motivation) zwar möglich, eine Hierarchisierung von Unterrichtszielen durch das eingeschränkte verfügbare Zeitbudget jedoch unvermeidlich ist. Demnach könnte sich mit einer stärkeren Betonung der selbstregulationsspezifischen Unterrichtsqualität die Schwerpunktsetzung im Unterricht verschieben. Möglich wären hierbei aufgrund individualisierender Elemente (z.B. Förderung der individuellen Bezugsnorm) der expliziten und impliziten Schülerförderung sowohl eine verstärkte Schülerorientierung im Fortbildungszeitraum als auch eine weniger starke Konzentration auf die Klassenführung. Schließlich wäre es in Anlehnung an Forschungsbefunde zur Expertiseentwicklung bei Lehrkräften durchaus auch möglich, dass die Implementation von expliziter und impliziter Selbstregulationsförderung im Unterricht die sonst berichtete Stabilität bzw. Automatisierung der Unterrichtsverläufe stört. Selbst für erfahrene Lehrkräfte könnte sich demnach die Komplexität des Unterrichts wieder vergrößern, da entwickelte Routinen, die bei erfahrenen Lehrkräften das Unterrichten erleichtern, durch die Integration neuer Elemente in den Unterricht unterbrochen werden (Berliner, 2001; Boshuizen, 2004; Bromme, 1992). So zeigte beispielsweise Bromme (1992), dass sich Experten bei für sie neuartigen Problemen und Aufgaben mehr Zeit für Überlegungen lassen als unerfahrene Lehrkräfte und Borko und Putnam (1995) stellen fest, dass eine fortbildungsbedingte Umgestaltung der Unterrichtspraxis von den Lehrkräften auch die Entwicklung neuer Strategien der Klassenführung erfordert.

Um prüfen zu können, welche Effekte Lehrkräftefortbildungen zur Förderung der Selbstregulation auf die allgemeine Unterrichtsqualität haben, ist es notwendig, zentrale Charakteristika allgemeiner Unterrichtsqualität zu identifizieren. Die vorliegende Studie baut hierbei auf das Modell der drei Basisdimensionen der Unterrichtsqualität nach Klieme et al. (2001) auf. Dieses wurde im Rahmen der TIMS-Videostudien entwickelt und integriert verschiedene Unterrichtsmerkmale, deren Wirksamkeit im Rahmen der Unterrichtsforschung theoretisch hergleitet und empirisch belegt wurde (vgl. Hattie, 2008; Helmke, 2004; Seidel & Shavelson, 2007). So gehen Klieme et al. (2001) davon aus, dass Unterricht gleichzeitig sowohl Kriterien der direkten Instruktion (Rosenshine & Furst, 1973) als auch Kriterien eines motivationsförderlichen (Deci & Ryan, 1985) und kognitiv aktivierenden (Mayer, 2004) Unterrichts erfüllen kann, da diese nur verschiedene, aber durchaus vereinbare Dimensionen von Unterricht darstellen. Im weiteren Verlauf des vorliegenden Beitrags sollen die drei Basisdimensionen unter dem Begriff allgemeine Unterrichtsqualität zusammengefasst werden.

Die erste der drei Basisdimensionen ist die Klassenführung, welche die Fähigkeit einer Lehrperson beschreibt, im Unterricht einen zügigen und gut organisierten Stundenablauf zu gewährleisten. Dazu zählt sowohl die schnelle und effektive Abhandlung von Disziplinproblemen als auch der notwendige Überblick über Klassenzimmerprozesse, um bei Schwierigkeiten und Problemen rasch und im Idealfall vorausblickend handeln zu können (Klieme, Lipowsky, Rakoczy & Ratzka, 2006; Ophardt & Thiel, 2008; Waldis, Grob, Pauli & Reusser, 2010).

Die zweite Dimension, die kognitive Aktivierung, umfasst alle Aktivitäten der Lehrkraft, die Schülerinnen und Schüler zu einer intensiven Beschäftigung mit dem Unterrichtsinhalt anregen (Klieme et al., 2006). Im Unterricht kann dies sowohl durch Aufgabenstellungen erreicht werden, die das Vorwissen der Schülerinnen und Schüler aktivieren, als auch durch die Anregung der Lernenden, selbständig ihre eigenen Lösungen zu überprüfen. Wichtig dabei ist, ein optimales Anspruchsniveau zu finden, so dass alle Schülerinnen und Schüler in der Auseinandersetzung mit den Lerninhalten gefordert werden, ohne dass es zu einer Überforderung kommt (Klieme et al. 2006; Kunter & Voss, 2011).

Die dritte Dimension ist die Schülerorientierung bzw. „konstruktive Unterstützung“ (vgl. Kunter & Voss, 2011), die eine Ausrichtung des Unterrichts auch nach den Bedürfnissen der Schülerinnen und Schüler beinhaltet, was sich insbesondere positiv auf die Motivation der Schülerinnen und Schüler auswirkt (Klieme et al., 2006). Um dies zu gewährleisten, müssen Aufgaben so implementiert werden, dass individuelle Lernprozesse angeregt, die Lernenden gleichzeitig jedoch bei Schwierigkeiten angeleitet und begleitet werden. Wichtig dabei ist das Wahrnehmen und Erkennen individueller Verständnisschwierigkeiten durch die Lehrkraft und ein geduldiger und respektvoller Umgang mit jeder einzelnen Schülerin und jedem einzelnen Schüler (Klieme et al., 2006; Kunter et al., 2005).

Ableitung der Fragestellung

Im Mittelpunkt der vorliegenden Studie steht die Frage, inwieweit die selbstregulationsbezogene Lehrkräftefortbildung „Lernen mit Plan“ aus Lehrer- und Schülersicht eine Veränderung der Unterrichtsqualität in Bezug auf die trainierten Bereiche (selbstregulationsspezifische Unterrichtsqualität) sowie weiterer Bereiche des Unterrichtsgeschehens (allgemeine Unterrichtsqualität) bewirkt hat. „Lernen mit Plan“ ist eine viertägige Lehrkräftefortbildung, die mithilfe eines Lehrermanuals geskriptete Unterrichtsstunden und Unterrichtsmaterialien bereit stellt. Zur Prüfung der Effekte von

„Lernen mit Plan“ wurde ein prä-post-follow-up-Design implementiert, wobei die follow-up-Erhebung sieben Wochen nach Ende der Umsetzung des Programms im Unterricht stattfand.

Folgende Forschungsfragen werden dabei überprüft:

Erstens wird untersucht, ob das Training dazu geführt hat, dass die selbstregulationsspezifische Unterrichtsqualität gesteigert wurde. Erwartet wird, dass das Training in der Experimentalgruppe im Vergleich zur Kontrollgruppe sowohl aus Lehrer- als auch aus Schülersicht zu einer unmittelbaren und langfristigen verstärkten Selbstregulationsförderung im Unterricht führte.

Zweitens wird untersucht, welche Effekte die Fortbildungsteilnahme im Hinblick auf Indikatoren der allgemeinen Unterrichtsqualität hatte. Hier lassen sich nicht ohne weiteres gerichtete Vorhersagen formulieren. Verschiedene Effekte sind denkbar (siehe Abschnitt 2): So könnten bspw. die individualisierenden Elemente der Schülerförderung zu positiven Transfereffekten auf weitere Aspekte der Schülerorientierung führen; denkbar ist auch, dass die starke Nutzung von manualisierten Stundenentwürfen und vorgeschlagenen Fördermaßnahmen – unter der Voraussetzung einer hohen Qualität dieser Elemente – zu einer besseren Strukturierung des Unterrichts sowie verstärkten Nutzung kognitiv aktivierender Elemente geführt hat. Aber auch negative Effekte durch die Unterbrechung eingespielter Unterrichtsroutinen sind möglich. In Anbetracht dieser potenziell konfligierenden Konsquenzen und aufgrund des Mangels an entsprechenden Forschungsbefunden wird zwar eine Veränderung der allgemeinen Unterrichtsqualität erwartet, es werden jedoch keine gerichteten Vorerwartungen bzgl. der unmittelbaren und langfristigen Entwicklung formuliert.

Methode

Das Projekt „Lernen mit Plan“

Im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Projekts „Lernen mit Plan“ wurde in einer Kooperation der Technischen Universität Darmstadt und der Universität Tübingen eine Lehrkräftefortbildung entwickelt und evaluiert. Da bisherige Selbstregulationstrainings fast ausschließlich an Gymnasien oder Grundschulen durchgeführt wurden, jedoch wenig über die Fördereffekte an Hauptschulen bekannt ist, richtete sich die Fortbildung „Lernen mit Plan“ an Haupt- und Werkrealschullehrerinnen und –lehrer[11] in Baden-Württemberg. Voraussetzung für die Anmeldung war, dass die betreffende Lehrkraft

[11] Die Werkrealschule ist eine weiterführende Schulform im baden-württembergischen Schulsystem, die im Anschluss an die Grundschule in einem durchgehenden Bildungsgang bis Klasse 10 zum Werkrealabschluss (MBA) führt und dabei einen Hauptschulabschluss in Klasse 9 und Klasse 10 ermöglicht. Das in der Werkrealschule gültige pädagogische Konzept wird gleichermaßen in der Hauptschule umgesetzt (vgl. http://www.kultusportal-bw.de/servlet/PB/menu/1247342/index.html).

zum Zeitpunkt der Fortbildung eine fünfte Klasse in Mathematik unterrichtete, um so eine bereits frühzeitige Selbstregulationsförderung zu ermöglichen. Die Gestaltung der Fortbildung war orientiert an aktuellen Befunden zur erfolgreichen Gestaltung von Lehrkräftefortbildungen sowohl im Allgemeinen (Darling-Hammond et al., 2009; Lipowsky, 2010) als auch spezifisch zur Förderung der Selbstregulation von Schülerinnen und Schülern (Dignath & Büttner, 2008).

Im Rahmen der Fortbildung wurde den teilnehmenden Lehrkräften in zwei zweitägigen Fortbildungsblöcken ein auf sechs Schulwochen angelegtes Programm zur Förderung des selbstregulierten Lernens der Schülerinnen und Schüler vermittelt, welches von den Lehrkräften anschließend in den Mathematikunterricht implementiert werden sollte. Die Fortbildung basierte dabei auf dem Prozessmodell des selbstregulierten Lernens nach Schmitz und Wiese (2006), das den Lernprozess in eine präaktionale, eine aktionale und eine postaktionale Phase unterteilt (vgl. Abbildung 1). Abbildung 1 zeigt zudem, dass die Schülerförderung sieben Elemente (Zielsetzung, Planung, Motivation, Volition bzw. Durchhaltevermögen, Umgang mit Fehlern und individuelle Bezugsnormorientierung und Reflexion) umfasste.

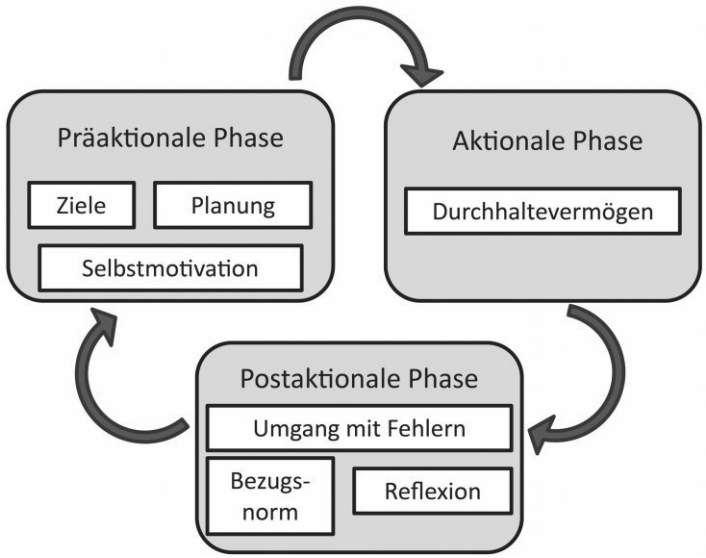

Abbildung 1: Modell der Vermittlung der Selbstregulation in der Fortbildung „Lernen mit Plan“ in Anlehnung an das Prozessmodell der Selbstregulation nach Schmitz und Wiese (2006).

Die Vermittlung dieser Elemente an die Schülerinnen und Schüler erfolgte dabei sowohl explizit als auch implizit. Im Rahmen von „Lernen mit Plan“ wurde die explizite Förderung

realisiert, indem das jeweilige Element des Selbstregulationskreislaufs in einer vorkonzipierten Unterrichtsstunde abgelöst von fachlichen Inhalten mit den Schülerinnen und Schülern erarbeitet wurde. Für die Umsetzung der expliziten Förderung erhielten die Lehrkräfte ein detailliertes Unterrichtsmanual, das zu jedem Element der Schülerförderung eine ausführliche Zusammenfassung des theoretischen Hintergrunds, der vermittelten Inhalte sowie der vorgesehenen Unterrichtsziele enthielt. Darüber hinaus wurden im Unterrichtsmanual für jede Förderwoche eine vorkonzipierte Unterrichtsstunde sowie alle dafür notwendigen Unterrichtsmaterialien bereitgestellt. So war beispielsweise für die erste Einheit der Schülerförderung zur Zielsetzung vorgesehen, mit den Schülerinnen und Schülern zunächst die Bedeutung von Zielen anhand eines Comics zu erarbeiten. Im Anschluss daran sollten den Schülerinnen und Schülern die Kriterien eines guten Ziels erklärt werden. Abschließend sollte jede Schülerin und jeder Schüler für sich ein schulisches und ein privates Ziel schriftlich formulieren. Eine Abschrift des schulischen Ziels sollte außerdem in einer „Schatztruhe“ von der Lehrkraft bis zum Ende der Schülerförderung verwahrt werden, um die Zielerreichung in der letzten Förderstunde zur Reflexion gemeinsam zu überprüfen.

Die implizite Förderung sollte im Sinne des *Scaffoldings* (Reinmann & Mandl, 2006) darüber erfolgen, dass die Lehrkraft die Selbstregulation der Schülerinnen und Schüler im regulären Mathematikunterricht durch gezielte Fragestellungen und Hinweise mit besonderem Schwerpunkt auf dem Element der jeweiligen expliziten Förderstunde anregte (vgl. Wirth, 2009). Vorschläge, wie die Inhalte der expliziten Förderstunden in den weiteren regulären Mathematikstunden implizit gefördert werden können, waren ebenfalls im Unterrichtsmanual enthalten. So wurde in jedem Abschnitt zur impliziten Förderung des jeweiligen Elements immer auch eine komplexe Mathematikaufgabe vorgestellt und eine Erklärung dazu gegeben, wie diese mit den Schülerinnen und Schülern selbstreguliert bearbeitet werden kann. Eine Auflistung weiterer Aufgaben aus den gängigen Mathematiklehrbüchern, die ebenso bearbeitet werden können, war ebenfalls im Manual enthalten. Darüber hinaus wurde den Lehrkräften beispielsweise im Abschnitt zur impliziten Förderung der Zielsetzung vorgeschlagen, die Schülerinnen und Schüler am Anfang jeder Schulwoche dazu aufzufordern, sich ein eigenes Ziel für diese Woche zu stecken (z.B. „Ich möchte mich diese Woche mindestens fünfmal im Unterricht melden.“). Sowohl das selbstregulierte Vorgehen bei der Bearbeitung von Aufgaben als auch die weiteren Vorschläge zur Vertiefung der Programminhalte wurden in der Fortbildung gemeinsam erarbeitet, so dass das Unterrichtsmanual vor allem dazu diente, die Lehrkräfte bei der Umsetzung zu entlasten und an die Inhalte der Fortbildung zu erinnern.

Stichprobe

Zum Programm „Lernen mit Plan“ hatten sich 96 Lehrkräfte der Haupt- und Werkrealschule in Baden-Württemberg mit ihren Klassen der fünften Jahrgangsstufe angemeldet. Die Teilnehmerinnen und Teilnehmer wurden randomisiert einer von zwei Experimentalgruppen bzw. der Wartekontrollgruppe, die im Fortbildungszeitraum regulär unterrichten sollte, zugewiesen. Da sich beide Experimentalgruppen nur darin unterschieden, dass eine der Gruppen zusätzlich im Verlauf der Fortbildung an zwei Nachmittagen an einem einstündigen Coachinggespräch zur Umsetzung der Fortbildungsinhalte teilnahm, ansonsten jedoch genau das gleiche Treatment erhielt, wurden für beide Gruppen positive Effekte der Fortbildung erwartet. Im Rahmen der für diesen Beitrag durchgeführten Analysen wurde daher nur nach der Zugehörigkeit zur Experimentalgruppe (EG) und der Wartekontrollgruppe (KG) unterschieden. Nach der Randomisierung befanden sich demnach 64 Lehrkräfte in der EG und 32 Lehrkräfte in der KG.

Aufgrund von Krankheit und organisatorischen Schwierigkeiten verringerte sich die Stichprobe zwischen dem Zeitpunkt der Zulosung und dem tatsächlichen Fortbildungsbeginn um 17 Personen, so dass zum ersten Messzeitpunkt noch 79 Lehrkräfte mit ihren 5. Klassen an der Datenerhebung teilnahmen. Davon gehörten 51 Lehrkräfte der EG und 28 Lehrkräfte der KG an.

Betrug die Rücklaufquote zum ersten Messzeitpunkt noch 100%, so lagen beim zweiten und dritten Messzeitpunkt jeweils nur 73 bzw. 72 Fragebögen vor, was einer Rücklaufquote von 91% entspricht. Dies ist darauf zurückzuführen, dass drei Lehrkräfte der Experimentalgruppe die Teilnahme nach dem ersten Fortbildungsblock aufgrund von beruflicher Veränderung oder mangelnder Akzeptanz des Programms abbrachen und ein Teilnehmer der Experimentalgruppe sowohl zum zweiten als auch dritten Messzeitpunkt keinen Fragebogen abgegeben hat. Darüber hinaus fehlten zum zweiten und dritten Messzeitpunkt jeweils zwei bzw. drei weitere Fragebögen von Lehrkräften, die jedoch an den anderen beiden Messzeitpunkten teilgenommen haben. Aufgrund inhaltlicher Überlegungen wurden die vier Lehrkräfte der Experimentalgruppe, von denen sowohl Angaben zum zweiten als auch zum dritten Messzeitpunkt fehlten, von den weiteren Analysen ausgeschlossen, wohingegen die Angaben von Lehrkräften, die nur zu einem Zeitpunkt keine Angaben gemacht hatten, mithilfe von multipler Imputation geschätzt wurden. Die den vorliegenden Analysen zugrunde liegende Stichprobe betrug insgesamt 75 Lehrkräfte.

Beschreibung der Lehrerstichprobe.

Das durchschnittliche Alter der 75 Lehrkräfte betrug 45.20 Jahre (*SD* = 11.73) und die durchschnittliche Anzahl der Dienstjahre betrug 17.14 Jahre (*SD* = 11.36). Einundfünfzig Personen waren weiblich und 24 männlich. Ungefähr die Hälfte der Lehrkräfte unterrichtete die Klasse, mit der sie an der Fortbildung teilnahm, nicht nur in Mathematik, sondern war außerdem Klassenlehrerin oder Klassenlehrer der betreffenden Klasse. Aufgrund der Randomisierung unterschieden sich die Gruppen in keinem der beschrieben Merkmale signifikant voneinander.

Beschreibung der Schülerstichprobe.

Im Rahmen von „Lernen mit Plan“ wurden die Angaben von insgesamt 1527 Schülerinnen und Schülern erfasst. Um eine größere Anzahl von Instrumenten einsetzen zu können, ohne die Befragten zu sehr zu beanspruchen, bearbeitete in jeder Klasse nur jeweils die Hälfte der Schülerinnen und Schüler den Fragebogenabschnitt zur Erfassung der Unterrichtsmerkmale. Für die Beurteilung der Fortbildungseffekte auf den Unterricht stehen daher pro Klasse die Angaben von durchschnittlich 9,5 Schülerinnen und Schülern zur Verfügung. Die Berechnung der Messzeitpunktdifferenz t2 - t1 basiert dabei auf den Angaben von insgesamt 690 (EG = 429, KG = 261) Schülerinnen und Schülern und die Schätzung der Messzeitpunktdifferenz t3 - t1 auf den Angaben von 689 (EG = 427, KG = 262) Schülerinnen und Schülern. Sowohl in der Experimentalgruppe als auch in der Kontrollgruppe betrug der Mädchenanteil ca. 47.7%. Das durchschnittliche Alter der Schülerinnen und Schüler betrug in der Experimentalgruppe 11.76 Jahre (*SD* = 0.74) und in der Kontrollgruppe 11.74 Jahre (*SD* = 0.74).

Die Erfassung des Unterrichts aus Lehrer- und Schülersicht

Die Umsetzung der Fortbildungsinhalte im Unterricht wurde durch Fragebögen auf Lehrer- und Schülerseite erfasst. Die Datenerhebung erfolgte zu drei Messzeitpunkten. Die erste Lehrer- und Schülerbefragung fand vor dem ersten Fortbildungsblock bzw. vor Beginn der Schülerförderung statt. Der zweite Messzeitpunkt lag in der Woche nach Abschluss der Schülerförderung und die dritte Messung zur Erfassung der Stabilität der Fortbildungseffekte erfolgte sieben Wochen nach dem zweiten Messzeitpunkt. Detaillierte Informationen zur internen Konsistenz der Skalen und Beispielitems sind Tabelle 1 zu entnehmen[12].

[12] Auf eine detaillierte Auskunft zur Herkunft der verwendeten Skalen wurde aus Platzgründen verzichtet. Eine genauere Dokumentation ist jedoch auf Anfrage bei der Erstautorin erhältlich.

Tabelle 1
Eigenschaften der verwendeten Lehrer- und Schülerskalen zum ersten Messzeitpunkt

Skalenname	*M*	*SD*	Reliabilität	Itemanzahl	Beispielitem
Lehrerfragebogen					
Allgemeine Unterrichtsqualität					
Klassenführung	2.36	0.74	.85	4 [a]	„In dieser Klasse wird der Unterricht kaum gestört.“
Effektive Zeitnutzung	2.73	0.54	.64	4 [a]	„Ich habe oft den Eindruck, dass im Unterricht in dieser Klasse viel Zeit vertrödelt wird.“ (rekodiert)
Kognitiv aktivierender Umgang mit Schülerbeiträgen	2.86	0.40	.60	5 [a, b]	„Ich lasse die Schüler(innen) auch einmal bewusst in die Irre laufen, bis sie sehen, dass etwas nicht stimmen kann.“
Allgemeine Unterstützung	3.69	0.38	.70	3 [a]	„Ich ermutige meine Schüler(innen), jederzeit zu fragen, wenn sie etwas nicht verstehen.“
Autonomieunterstützung	3.26	0.32	.66	7 [a, b]	„Bei mir dürfen die Schüler(innen) zum Lösen von Aufgaben ihre eigenen Strategien einsetzen.“
Leistungsdifferenzierung	2.44	0.62	.68	4 [a]	„Leistungsstarken Schüler(inne)n gebe ich Extraaufgaben, durch die sie wirklich gefordert werden.“
Selbstregulationsspezifische Unterrichtsqualität					
Unterstützung der Zielsetzung	3.01	0.40	.83	10 [a, b]	„Während des Mathematikunterrichts...ermuntere ich meine Schüler(innen), sich selbst Ziele zu setzen.“
Unterstützung der Planung	2.39	0.66	.68	2 [b]	„Während des Mathematikunterrichts...achte ich darauf, dass meine Schüler(innen) sich vor der Bearbeitung einer komplexen Aufgabe einen Plan machen.“
Unterstützung der Konzentration	2.64	0.40	.74	6 [b]	„Während des Mathematikunterrichts...zeige ich meinen Schüler(inne)n, was sie gegen abschweifende Gedanken selbst tun können.“
Unterstützung der Motivation	2.37	0.49	.78	4 [b]	„Während des Mathematikunterrichts...ermutige ich meine Schüler dazu, Selbstmotivationsstrategien anzuwenden.“
Unterstützung beim Umgang mit Fehlern	3.29	0.43	.57	2 [b]	„Während des Mathematikunterrichts...helfe ich meinen Schüler(innen)n dabei, mit Fehlern konstruktiv umzugehen.“
Förderung der individuellen Bezugsnorm bei der Leistungsrückmeldung	2.70	0.41	.62	7 [a, b]	„Nach Klassenarbeiten…schreibe ich bei der Korrektur von Klassenarbeiten meinen Schüler(inne)n unter die Note, was sie gut gemacht haben, und was sie noch verbessern können.“

Fortsetzung von Tabelle 1

Unterstützung der Selbstregulation	2.62	0.60	.70	3 [b]	„Während des Mathematikunterrichts...unterstütze ich meine Schüler(innen) bei der Bearbeitung komplexer Aufgaben dabei, diese selbstreguliert zu bearbeiten.“
Schülerfragebogen					
Allgemeine Unterrichtsqualität					
Disziplinäres Klima	2.70	0.71	.76	4 [c]	„Bei unserem Mathelehrer… können wir ungestört arbeiten.“
Effektive Zeitnutzung	2.44	0.77	.59	3 [c]	„Bei unserem Mathelehrer… wird im Unterricht viel Zeit vertrödelt.“ (rekodiert)
Diskursive Behandlung von Schülerbeiträgen	3.06	0.57	.73	6 [c]	„Im Matheunterricht sagt der Mathelehrer öfters nicht gleich, ob eine Antwort falsch oder richtig ist.“
Allgemeine Unterstützung	3.31	0.64	.78	4 [c]	„Unser Mathelehrer…ermutigt uns zu fragen, wenn wir etwas nicht verstehen.“
Autonomieunterstützung	3.07	0.61	.74	5 [c]	„Bei meinem Mathelehrer kann ich zum Lösen schwieriger Aufgaben meine eigenen Strategien einsetzen.“
Selbstregulationsspezifische Unterrichtsqualität					
Zielorientierung	2.99	0.66	.75	5 [c]	„Unser Mathelehrer kann uns gut seine Ziele erklären.“
Fehlerkultur	3.10	0.67	.71	4 [c]	„Wenn ich im Matheunterricht etwas falsch mache, erhalte ich die Möglichkeit, mich zu korrigieren oder noch einmal anzufangen.“

Anmerkungen. Lehrerangaben: n = 71 - 75; Schülerangaben: n = 523 – 607.
[a] Antwortformat 1: 1 (*trifft nicht zu)*, 2 (*trifft eher nicht zu)*, 3 (*trifft eher zu)*, 4 (*trifft zu).*
[b] Antwortformat 2: 1 ((*fast) nie)* , 2 (*selten)*, 3 (*oft)*, 4 ((*fast) immer).*
[c] Antwortformat 3: 1 (*stimmt gar nicht)*, 2 (*stimmt eher nicht)*, 3(*stimmt eher)*, 4 (*stimmt genau).*

Selbstregulationsspezifische Unterrichtsqualität.

Die Skalen zur Erfassung der Selbstregulationsförderung im Unterricht basieren auf mehreren Studien zur Verbesserung der Selbstregulation bei Schülerinnen und Schülern wie z.B. PROSA (Komorek, Bruder, Collet & Schmitz, 2006) und SELVES (Otto, 2007) und wurden durch Ihringer (in Vorbereitung) auf die Bedürfnisse von „Lernen mit Plan“ angepasst.

Im Lehrerfragebogen wurden zur Erfassung der Umsetzung der behandelten Fortbildungsinhalte die Skalen Unterstützung der Zielsetzung, Unterstützung der Planung, Unterstützung der Motivation, Unterstützung der Konzentration, Förderung der individuellen Bezugsnorm bei der Leistungsrückmeldung und Unterstützung beim Umgang mit Fehlern eingesetzt. Darüber hinaus wurde die Skala Unterstützung der Selbstregulation eingesetzt, welche die generelle Selbstregulationsförderung im Mathematikunterricht erfasst. Die Skala Unterstützung der Reflexion, die 2 Items umfasste (Beispielitem: „Während des Mathematikunterrichts ... unterstütze ich meine Schüler(innen) darin, ihre Vorgehensweise oder ihr Ziel gegebenenfalls eigenständig zu modifizieren.“), wurde aufgrund mangelnder interner Konsistenz (Cronbachs $\alpha = .43$) von den Analysen ausgeschlossen.

Da im Rahmen der Schülerbefragung auch zahlreiche Informationen zum Hintergrund der Schülerinnen und Schüler und deren eigener Selbstregulation und Motivation erfasst wurden, konnte aus Platzgründen keine vollständig Parallelisierung der im Schüler- und Lehrerfragebogen eingesetzten Skalen zur Erfassung der Unterrichtsqualität realisiert werden. Die Erfassung der selbstregulationsspezifischen Unterrichtsqualität erfolgte daher im Schülerfragebogen nur durch die Skalen *Zielorientierung* und *Fehlerkultur*.

Wie Tabelle 1 zu entnehmen ist, besaßen alle verwendeten Skalen zur Erfassung der selbstregulationsspezifischen Unterrichtsqualität ein vierstufiges Antwortformat, wobei eine höhere Einschätzung auf diesen Skalen einer höheren Ausprägung des jeweiligen Konstrukts entspricht. Mit Ausnahme der Skala *Unterstützung beim Umgang mit Fehlern*, welche zum ersten Messzeitpunkt lediglich eine interne Konsistenz (Cronbachs Alpha) von $\alpha = .57$ aufwies, sind die internen Konsistenzen der Skalen zur Erfassung der selbstregulationsspezifischen Unterrichtsqualität als ausreichend bis gut ($.60 \leq \alpha \leq .87$) einzustufen.

Allgemeine Unterrichtsqualität.

Die Entwicklung der Skalen zur allgemeinen Unterrichtsqualität erfolgte aufbauend auf einer Reihe unterschiedlicher Skalen aus mehreren nationalen und internationalen Studien wie

bspw. COACTIV (Baumert et al., 2009), BIJU (Baumert, Gruehn, Heyn, Köller & Schnabel, 1997), PYTHAGORAS (Rakoczy, Buff & Lipowsky, 2005), DESI (Wagner, Helmke & Rösner, 2009), PALMA (Pekrun et al., 2002) und PROSA (Komorek et al., 2006). Gemäß der Fragestellung der vorliegenden Studie wurde die Unterrichtsgestaltung der Lehrkräfte hinsichtlich der drei zuvor beschriebenen Basisdimensionen der Unterrichtsqualität (Klassenführung, kognitive Aktivierung und Schülerorientierung; Klieme et al., 2001) durch Lehrer- und Schülerauskünfte erfasst.

In den Lehrerfragebögen wurde die Dimension Klassenführung durch die Skalen *Klassenführung* und *effektive Zeitnutzung* repräsentiert. Die Dimension kognitive Aktivierung wurde durch die Skala *kognitiv aktivierender Umgang mit Schülerbeiträgen* erfasst. Die Dimension Schülerorientierung wurde durch die drei Skalen *Leistungsdifferenzierung, Autonomieunterstützung* und *allgemeine Unterstützung* erfasst.

Im Schülerfragebogen wurden zur Erfassung der allgemeinen Unterrichtsqualität die Skalen disziplinäres Klima, effektive Zeitnutzung, diskursive Behandlung von Schülerbeiträgen, Autonomieunterstützung und allgemeine Unterstützung eingesetzt, welche inhaltlich den Skalen im Lehrerfragebogen entsprachen.

Wie Tabelle 1 zu entnehmen ist, besaßen alle verwendeten Skalen ein vierstufiges Antwortformat und zeigten mit Ausnahme der Skalen *kognitiv aktivierender Umgang mit Schülerbeiträgen* (Lehrerfragebogen), welche zum dritten Messzeitpunkt ein Cronbachs Alpha von .56 aufwies, und der Skala *effektive Zeitnutzung* (Schülerfragebogen), die zum ersten Messzeitpunkt ein Cronbachs Alpha von α =.59 aufwies, eine ausreichende bis gute interne Konsistenz ($.60 \leq \alpha \leq .89$).

Statistische Analysen

In der vorliegenden Studie wurden basierend auf Lehrer- und Schülerauskünften verschiedene allgemeine sowie selbstregulationsspezifische Unterrichtsmerkmale untersucht Dabei wurden jeweils Veränderungen zwischen verschiedenen Messzeitpunkten (unmittelbare bzw. langfristige Treatment-Effekte) berücksichtigt. Um eine Alpha-Fehler-Kumulierung bei mehreren geplanten (konfirmatorischen) Gruppenvergleichen zu vermeiden, wurden zunächst für die Überprüfung der Hypothesen relevante, skalenspezifische Differenzwerte als Maß für die Veränderung zwischen den jeweiligen Messzeitpunkten (t2 - t1, t3 - t1) gebildet und dann gemäß der Annahmen über ihre Veränderung durch die Intervention zu zwei Globalmaßen (Schochet, 2008) für die allgemeine bzw. die selbstregulationsspezifische Unterrichtsqualität zusammengefasst. Diese vier neu gebildeten abhängigen Variablen für die jeweilige

Perspektive—für jede der beiden Messzeitpunktdifferenzen jeweils zwei Globalmaße—wurden dann anhand von für multiple Signifikanztests adjustierten *p*-Werten auf signifikante Mittelwertunterschiede zwischen Kontroll- und Experimentalgruppe getestet.

Für signifikante Unterschiede eines Komposits wurden anschließend unadjustierte Tests bezüglich der Differenzwerte der einzelnen Skalen des Komposits berechnet, um Veränderungen auf der Einzelskalenebene statistisch abzusichern (Schochet, 2008). Bei nicht signifikanten Ergebnissen eines Komposits wurden stattdessen exploratorische Analysen für die Teilskalen durchgeführt, um auszuschließen, dass das Ausbleiben eines Effektes auf gegenläufige Effekte der Teilskalen des Komposits zurückzuführen ist. Alle berichteten *p*-Werte wurden dabei im Rahmen einer zweiseitigen Signifikanztestung ermittelt, sind aber gemäß der formulierten Hypothesen im Fall des Globalmaßes und der Teilskalen der *selbstregulationsspezifischen Unterrichtsqualität* einseitig zu interpretieren.

Zur Analyse der Lehrerangaben wurden mithilfe der Software IVEware (Raghunathan, Solenberger & Hoewyk, 2002) fehlende Werte anhand der Hintergrundmerkmale *Geschlecht*, *Alter*, *Gewissenhaftigkeit*, *Unterrichtserfahrung in Mathematik* [in Jahren], *Klassenlehrer der Klasse* und *Wochenstunden Mathematikunterricht in der Fortbildungsklasse* mittels multipler Imputation (insgesamt fünf Werte) ersetzt. Die durchgeführten Analysen beruhen daher auf einem Datensatz, bei dem die jeweils fünf imputierten Werte (pro Skala und Fall) durch Mittelwertbildung zusammengefasst wurden. Die Mittelwerte der Globalmaße wurden mithilfe der im Statistikpaket SAS (SAS Institute Inc., 2004) implementierten Prozedur MULTTEST anhand von Bootstrap-Replikationen auf signifikante Unterschiede zwischen den Gruppen getestet. Die verwendete MULTTEST Prozedur zeichnet sich dabei im Vergleich zur Šidák- oder Bonferroni-Korrektur durch eine höhere Teststärke aus (Westfall & Young, 1993), was insbesondere bei geringen Stichprobenumfängen von Vorteil ist.

Hinsichtlich der Analyse der Schülerangaben ist zu berücksichtigen, dass die Implementation der Trainingsinhalte auf Klassenebene erfolgte und darüber hinaus davon auszugehen ist, dass die Angaben der Schülerinnen und Schüler innerhalb von Klassen nicht unabhängig voneinander sind. Es liegt daher eine hierarchische Datenstruktur vor, wobei Schülerinnen und Schüler in Klassen gruppiert sind, die wiederum entweder der Experimental- oder der Kontrollgruppe zuzuordnen sind. Bei der Analyse der Schülerdaten muss demnach die Mehrebenenstruktur der Daten berücksichtigt werden. Dies ist jedoch im Rahmen der zur Analyse der Lehrerangaben verwendeten SAS-Prozedur MULTTEST nicht möglich. Differenzielle Veränderungen zwischen zwei Messzeitpunkten (Unterschiede der Differenzenwerte zwischen Experimental- vs. Kontrollgruppe) wurden daher anhand von

Pfadmodellen mit dem Programm Mplus (Muthén & Muthén, 1998-2010) ermittelt. Die auf die vier Globalmaße bezogenen Interventionseffekte wurden anhand einer Aggregation der betreffenden merkmalsspezifischen, differentiellen Veränderungen zu jeweils einem Mittelwert pro Intervall und Gruppe modelliert. Fehlende Werte wurden dabei mithilfe des FIML-Verfahrens (Full Information Maximum Likelihood) geschätzt. Die Adjustierung der *p*-Werte dieser Globalmaße für multiple Signifikanztests erfolgte mithilfe der Šidák-Korrektur (Šidàk, 1967) $p_{\text{adjustiert}} = 1 - (1 - p_{\text{unadjustiert}})^k$, wobei k der Anzahl der Tests entspricht.

Ergebnisse

Veränderung der selbstregulationsspezifischen bzw. allgemeinen Unterrichtsqualität aus Lehrersicht.

Zur Überprüfung der Hypothesen wurden nach dem oben beschriebenen Verfahren auf der Basis der skalenspezifischen Differenzwerte zwischen den jeweiligen Messzeitpunkten (t2 - t1, t3 - t1) die Globalmaße selbstregulationsspezifische Unterrichtsqualität und allgemeine Unterrichtsqualität gebildet und auf Unterschiede zwischen den Gruppen getestet. Wie Tabelle 2 entnommen werden kann, erwiesen sich unter Berücksichtigung der für multiple Tests adjustierten *p*-Werte beide Messzeitpunktdifferenzen des Globalmaßes *selbstregulationsspezifische Unterrichtsqualität* (t2 - t1, t3 - t1) im Vergleich zur Kontrollgruppe als statistisch signifikant.

Tabelle 2

Messzeitpunktdifferenzen (t2 - t1, t3 - t1) der beiden Globalmaße allgemeine Unterrichtsqualität und selbstregulationsspezifische Unterrichtsqualität getrennt nach Gruppen aus Lehrer- und Schülerperspektive

Globalmaße für Veränderungen (Differenzwerte)	Gruppe	*M*	*SD*	*p*-Wert (adjustiert)
Selbstregulationsspezifische Unterrichtsqualität t2 - t1	KG EG	0.00/*-0.19* 0.24/*-0.04*	0.22/*0.56* 0.28/*0.65*	.001 [a]/*.016* [b]
Selbstregulationsspezifische Unterrichtsqualität t3 - t1	KG EG	0.01/*-0.40* 0.16/*-0.10*	0.24/*0.67* 0.29/*0.65*	.073 [a]/*.000* [b]
Allgemeine Unterrichtsqualität t2 - t1	KG EG	-0.02/*-0.15* 0.04/*-0.06*	0.21/*0.41* 0.20/*0.43*	.561 [a]/*.043* [b]
Allgemeine Unterrichtsqualität t3 - t1	KG EG	-0.01/*-0.26* 0.01/*-0.10*	0.27/*0.47* 0.24/*0.46*	.997 [a]/*.000* [b]

Anmerkungen. Lehrerstichprobe: KG = Kontrollgruppe (n = 28), EG = Experimentalgruppe (n = 47); Schülerstichprobe: KG = Kontrollgruppe (n = 261 - 262), EG = Experimentalgruppe (n = 429 - 427). Die Angaben für die Schülerstichprobe sind kursiv geschrieben.
[a]. Signifikanztest mit Adjustierung für multiple Tests.
[b]. Signifikanztest mit Šidák-Korrektur für multiple Tests.

Dabei ist auch die Messzeitpunktdifferenz von t3 - t1 statistisch signifikant, da im Falle der selbstregulationsspezifischen Unterrichtsqualität der in Tabelle 2 für adjustierte zweiseitige Tests berichtete *p*-Wert in Erwartung eines Anstiegs als einseitiger Test zu interpretieren ist. Die Annahme einer Veränderung der selbstregulationsspezifischen Unterrichtsqualität im Verlauf der Fortbildung kann daher ebenso bestätigt werden wie die erwartete Stabilität dieser Effekte zum dritten Messzeitpunkt. Wie Tabelle 2 zu entnehmen ist, konnte für das Globalmaß *allgemeine Unterrichtsqualität* im Vergleich zur Kontrollgruppe weder für das Intervall t2 - t1 noch für das Intervall t3 - t1 eine signifikante Veränderung festgestellt werden.

Um zu überprüfen, ob sich differentielle Effekte für die Teilskalen der gebildeten Globalmaße ergeben, wurden anschließend skalenspezifische Differenzwerte auf Unterschiede zwischen den Gruppen untersucht. Da sich bei Untersuchung der Globalmaße die stärkste gruppenspezifische Veränderung zwischen dem ersten und zweiten Messzeitpunkt ergeben hat, beschränken sich die folgenden Analysen auf dieses Intervall. Wie Tabelle 3 zu entnehmen ist, zeigten sich hierbei für die Experimentalgruppe für alle Teilskalen außer *Förderung der individuellen Bezugsnorm bei der Leistungsrückmeldung* und *Unterstützung beim Umgang mit Fehlern*, die beide in der fünften Woche der Schülerförderung trainiert wurden, im Vergleich zur Kontrollgruppe statistisch signifikante Zuwächse mit einer Effektstärke[13] (Cohens *d*) von $0.52 \leq d \leq 0.65$. Dabei ist auch der Unterschied hinsichtlich der Skala *Unterstützung der Planung* mit $p = .053$ aufgrund der einseitigen Formulierung der Annahmen als signifikant anzusehen. Um zu überprüfen, ob gegenläufige Veränderungen auf den einzelnen in das Globalmaß *allgemeine Unterrichtsqualität* eingegangenen Skalen zu einem Ausbleiben eines Effektes geführt haben, wurde eine separate Testung der skalenspezifischen Differenzwerte der betreffenden Teilskalen vorgenommen. Da es sich hierbei um eine exploratorische Analyse handelt, werden für diese Analysen die für multiple Tests adjustierten *p*-Werte interpretiert. Wie Tabelle 3 entnommen werden kann, wies keine der Teilskalen der *allgemeinen Unterrichtsqualität* für die Experimentalgruppe statistisch signifikante positive oder negative Zuwächse auf. Es kann daher davon ausgegangen werden, dass die Fortbildung aus Sicht der teilnehmenden Lehrkräfte weder zu einer Verbesserung noch zu einer Verschlechterung im Bereich der erfassten Merkmale der allgemeinen Unterrichtsqualität geführt hat.

[13] Zur Berechnung der Effektstärke Cohens *d* wurde der Effekt der Gruppenzugehörigkeit unter Einbezug der Standardabweichung der betreffenden Skala zum ersten Messzeitpunkt standardisiert.

Tabelle 3

Unstandardisierte Differenzwerte der Skalen zur Selbstregulationsförderung im Unterricht und der allgemeinen Unterrichtsqualität zwischen Messzeitpunkt 1 und 2 aus Lehrerperspektive

Selbstregulationsförderung im Unterricht (Differenzwert)	Gruppe	*M*	*SD*	*p*-Wert	Cohen‘s *d*
Unterstützung der Zielsetzung	KG	-0.01	0.36	.006 [a]	0.56
	EG	0.22	0.32		
Unterstützung der Planung	KG	-0.02	0.74	.053 [a]	0.55
	EG	0.35	0.71		
Unterstützung der Konzentration	KG	0.02	0.33	.020 [a]	0.52
	EG	0.23	0.37		
Unterstützung der Motivation	KG	0.08	0.43	.016 [a]	0.65
	EG	0.40	0.59		
Förderung der Individuellen Bezugsnorm bei der Leistungsrückmeldung	KG	-0.08	0.40	.291 [a]	-
	EG	0.03	0.43		
Unterstützung beim Umgang mit Fehlern	KG	0.03	0.35	.337 [a]	-
	EG	0.14	0.51		
Unterstützung der Selbstregulation	KG	-0.01	0.43	.014 [a]	0.56
	EG	0.22	0.59		
Allgemeine Unterrichtsqualität (Differenzwert)					
Klassenführung	KG	-0.08	0.44	.998 [b]	-
	EG	-0.03	0.52		
Effektive Zeitnutzung	KG	0.08	0.45	.957 [b]	-
	EG	-0.02	0.54		
Kognitiv aktivierender Umgang mit Schülerbeiträgen	KG	0.01	0.35	.993 [b]	-
	EG	0.06	0.40		
Leistungsdifferenzierung	KG	0.06	0.44	.952 [b]	-
	EG	0.14	0.45		
Allgemeine Unterstützung	KG	-0.15	0.44	.807 [b]	-
	EG	-0.03	0.39		
Autonomieunterstützung	KG	-0.06	0.30	.134 [b]	-
	EG	0.10	0.27		

Anmerkungen. KG = Kontrollgruppe (n = 28), EG = Experimentalgruppe (n = 47).
[a]. Signifikanztest ohne Adjustierung für multiple Tests.
[b]. Signifikanztest mit Adjustierung für multiple Tests.

Veränderung der selbstregulationsspezifischen bzw. allgemeinen Unterrichtsqualität aus Schülersicht.

Tabelle 2 zeigt, dass sich beide Messzeitpunktdifferenzen des Globalmaßes *selbstregulationsspezifische Unterrichtsqualität* (t2 - t1, t3 - t1) der Experimental- im Vergleich zur Kontrollgruppe als statistisch signifikant erweisen. Die Annahme einer differentiellen Veränderung der selbstregulationsspezifischen Unterrichtsqualität in Abhängigkeit von der Gruppenzugehörigkeit kann daher ebenso bestätigt werden wie die erwartete Stabilität dieser Effekte zum dritten Messzeitpunkt. Im Gegensatz zum Bericht der Lehrkräfte unterschieden sich bei Analyse der Schülerberichte jedoch auch die untersuchten Messzeitpunktdifferenzen des Globalmaßes *allgemeine Unterrichtsqualität* (t2 - t1, t3 - t1) signifikant zwischen der Kontroll- und der Experimentalgruppe. Die signifikanten Unterschiede zwischen den beiden Untersuchungsgruppen ergaben sich vor allem dadurch, dass bei den Schülerinnen und Schülern der Experimentalgruppe ein geringerer Abfall der untersuchten Unterrichtsmerkmale zu verzeichnen war als in der Kontrollgruppe.

Zur Überprüfung differentieller Effekte für die Teilskalen der gebildeten Globalmaße, wurden anschließend skalenspezifische Differenzwerte auf Unterschiede zwischen den Gruppen untersucht. Da sich bei der Untersuchung der Globalmaße die stärkste gruppenspezifische Veränderung zwischen dem ersten und dritten Messzeitpunkt ergeben hat, beschränken sich die folgenden Analysen auf dieses Intervall. Wie Tabelle 4 zu entnehmen ist, zeigten sich dabei für die beiden selbstregulationsspezifischen Teilskalen *Zielorientierung* und *Fehlerkultur* im Vergleich zur Kontrollgruppe statistisch signifikante Unterschiede, die einer Effektstärke[14] von $0.42 \leq d \leq 0.49$ entsprechen. Dabei weist die Experimentalgruppe eine hypothesenkonform höhere Ausprägung auf den beiden Teilskalen auf als die Kontrollgruppe. In Bezug auf die Überprüfung der in das Globalmaß *allgemeine Unterrichtsqualität* eingegangenen Teilskalen wurde eine separate Testung der skalenspezifischen Differenzwerte vorgenommen. Dabei zeigten sich für alle Teilskalen außer für die Skala *effektive Zeitnutzung* signifikante Unterschiede zwischen den beiden Untersuchungsgruppen, wobei die ermittelten Effektstärken in einem Bereich von $0.25 \leq d \leq 0.39$ liegen. Wie im Fall der Globalmaße und der Teilskalen für die *selbstregulationsspezifische Unterrichtsqualität* ergaben sich die Unterschiede auf den

[14] Zur Berechnung der Effektstärke Cohens *d* wurde der Effekt der Gruppenzugehörigkeit unter Einbezug der Gesamtvarianz der betreffenden Skala zum ersten Messzeitpunkt standardisiert. Auf eine Standardisierung unter Einbezug der Varianz auf Klassenebene wurde dabei verzichtet, da die Effekte aufgrund der geringen Interklassenkorrelationen ($0.04 \leq \rho \leq 0.17$) wesentlich höher ausgefallen wären und einen Wertebereich von $0.62 \leq d \leq 1.72$ eingenommen hätten

betreffenden Teilskalen der allgemeinen Unterrichtsqualität dadurch, dass bei der Experimentalgruppe ein wesentlich geringeres Absinken der jeweiligen Werte zu finden war als bei der Kontrollgruppe. Bei den Teilskalen *disziplinäres Klima* und *effektive Zeitnutzung* fanden sich für die Experimentalgruppe aus Schülersicht leichte, aber nicht statistisch signifikante Verbesserungen.

Diskussion

Im Programm „Lernen mit Plan“ wurde eine Fortbildung realisiert, welche die explizite und implizite Selbstregulationsförderung durch die Lehrkräfte im Unterricht verbessern sollte. In Übereinstimmung mit den forschungsleitenden Annahmen ergaben sich für die Lehrkräfte der Experimentalgruppe auf fast allen erhobenen Aspekten der selbstregulationsspezifischen Unterrichtsqualität im Vergleich zur Kontrollgruppe signifikante Verbesserungen, die sich auch zum dritten Messzeitpunkt noch nachweisen ließen. Auch bei den Schülerinnen und Schülern ließ sich dieser positive Effekt der Fortbildungsteilnahme auf die erhobenen Aspekte der selbstregulationsspezifischen Unterrichtsqualität feststellen, die sowohl zum zweiten als auch dritten Messzeitpunkt signifikant höher ausgeprägt war als die der Kontrollgruppe. Das Ausmaß der aus Lehrer- und Schülersicht wahrgenommenen Fortbildungseffekte auf die selbstregulationsspezifische Unterrichtsqualität kann hierbei mit Effektstärken von $0.42 \leq d \leq 0.65$ als bedeutsam angesehen werden.

Weniger einheitlich sind die Befunde zur Auswirkung der Fortbildung auf die allgemeine Unterrichtsqualität. So unterschieden sich die Angaben der Lehrkräfte sowohl direkt nach Abschluss der Schülerförderung als auch sieben Wochen später auf keiner der Dimensionen der allgemeinen Unterrichtsqualität von den Angaben der Kontrollgruppe. Diese Befunde konnten auch bei Betrachtung auf Ebene der Einzelskalen zur Erfassung der Dimensionen Klassenführung, kognitive Aktivierung und Schülerorientierung bestätigt werden. Im Gegensatz dazu deuten die Angaben der Schülerinnen und Schüler der Experimentalgruppe im Vergleich zur Kontrollgruppe auf positive Effekte der Fortbildungsteilnahme auf die erhobenen Aspekte der allgemeinen Unterrichtsqualität hin, die mit Effektstärken von $0.25 \leq d \leq 0.39$ ebenfalls als bedeutsam angesehen werden können.

Unter gleichzeitiger Berücksichtigung der Lehrer- und Schülerperspektive lässt sich daher folgern, dass die Teilnahme an der Fortbildung „Lernen mit Plan“ einen positiven Effekt auf die Selbstregulationsförderung im Unterricht hatte und keine Verschlechterung bzw. möglicherweise sogar eine leichte Verbesserung der allgemeinen Unterrichtsqualität eingetreten ist. Zwar deuten die Angaben der Schülerinnen und Schüler sowohl in der

Experimental- als auch in der Kontrollgruppe gleichermaßen auf eine Verschlechterung sowohl der selbstregulationsspezifischen als auch der allgemeinen Unterrichtsqualität hin. Das—verglichen mit der Kontrollgruppe—geringere Absinken der selbstregulationsspezifischen sowie der allgemeinen Unterrichtsqualität in der Experimentalgruppe kann jedoch als positiver Effekt der Fortbildungsteilnahme auf die wahrgenommene Unterrichtsqualität interpretiert werden.

Hinsichtlich der Wirksamkeit von Lehrkräftefortbildungen zur Förderung der Selbstregulation lässt sich aus den Befunden des vorliegenden Beitrags schließen, dass es von maßgeblicher Bedeutung ist, im Rahmen der Fortbildungsevaluation nicht nur Effekte auf fokale Trainingsinhalte, sondern auch die Auswirkungen auf die allgemeine Unterrichtsqualität zu erheben. So geht beispielsweise die bei Schülerinnen und Schülern ermittelte Verbesserung des Strategiewissens häufig nicht mit einem demzufolge erwartbaren Leistungszuwachs einher (De Corte, Verschaffel & Van de Ven, 2001; Schreblowski & Hasselhorn, 2001). Des Weiteren wurde bisher nur unzureichend überprüft, ob die Komponenten des selbstregulierten Lernens die tatsächliche Ursache für die Wirksamkeit der jeweiligen Programme sind (Spörer & Glaser, 2010). Die im Rahmen von „Lernen mit Plan“ ermittelte Veränderung nicht nur der fokalen Trainingsinhalte, sondern auch der allgemeinen Unterrichtsqualität könnte hierbei ein Hinweis zur weiteren Aufklärung potenzieller Wirkmechanismen von selbstregulationsfördernden Fortbildungen sein. So könnte es durchaus sein, dass auch in anderen Lehrkräftefortbildungen zur Förderung der Selbstregulation die allgemeine Unterrichtsqualität beeinflusst wurde. Zwar zeigten sich bei „Lernen mit Plan“ im Vergleich zur Kontrollgruppe positive Effekte auf die allgemeine Unterrichtsqualität, jedoch könnten in Anlehnung an Kline, Deshler und Schumaker (1992) weniger vorstrukturierte Fortbildungen möglicherweise auch mit negativen Effekten auf die allgemeine Unterrichtsqualität einhergehen. Die damit einhergehende Verschlechterung der allgemeinen Unterrichtsqualität könnte dann trotz einer verbesserten Selbstregulationsförderung im Unterricht zu einem Ausbleiben des erwartbaren Leistungszuwachses der Schülerinnen und Schüler führen.

Insbesondere im Hinblick auf die unterschiedlichen Befunde bezüglich der Fortbildungseffekte aus Lehrer- und Schülersicht zeigt sich die zentrale Bedeutung eines multiperspektivischen (z.B. Lehrer- und Schülerangaben) und multidimensionalen (z.B. unterschiedliche Aspekte der Unterrichtsqualität) Ansatzes. So beschränkt sich die Evaluation der meisten Fortbildungen zur Förderung der Selbstregulation auf die Urteile von Schülerinnen und Schülern, wohingegen die Einschätzung der Lehrkräfte aufgrund von

geringen Stichprobengrößen nicht systematisch berücksichtigt werden kann. Insbesondere in Anbetracht eines Mangels an Befunden zur Übereinstimmung von Lehrer- und Schülerwahrnehmungen der Unterrichtsqualität im Rahmen von Interventionsstudien wäre es jedoch wichtig, beide Perspektiven systematisch zu erfassen, um hier weiterführende Erkenntnisse zu spezifischen Einflüssen auf die Lehrer- und Schülerwahrnehmung zu gewinnen (Clausen, 2002; Wagner, 2008). Eine mögliche Erklärung für die im vorliegenden Beitrag gefundenen Unterschiede zwischen beiden Wahrnehmungsperspektiven könnte sein, dass neben einer Verbesserung der selbstregulationsspezifischen Unterrichtsqualität zwar eine Veränderung der allgemeinen Unterrichtsqualität stattgefunden hat, diese von den Lehrkräften aber nicht bemerkt wurde. So waren die Lehrkräfte zum Zeitpunkt der Fortbildung stark auf die Umsetzung der Fortbildungsinhalte konzentriert, was dazu geführt haben könnte, dass sie eine Veränderung der Klassenführung, der kognitiven Aktivierung und der Schülerorientierung im Gegensatz zu den Schülerinnen und Schülern nicht bewusst wahrgenommen haben. In Anlehnung an Befunde zur unterschiedlichen Differenziertheit der Lehrer- und Schülerwahrnehmung könnte aber auch argumentiert werden, dass von den Schülerinnen und Schülern aufgrund einer eher globalen Unterrichtswahrnehmung (Kunter et al., 2005)—trotz eines ausschließlich spezifischen Fortbildungseffekts auf die selbstregulationsspezifische Unterrichtsqualität—auch Veränderungen hinsichtlich allgemeiner Unterrichtsmerkmale wahrgenommen wurden. Dagegen spricht jedoch die unterschiedliche Stärke der Veränderungen der selbstregulationsspezifischen und allgemeinen Unterrichtsqualität aus der Schülerperspektive.

Stärken, Schwächen und Ausblick

Der vorliegende Beitrag ist unseres Wissens der erste Ansatz, im Rahmen von Lehrkräftefortbildungen zur Förderung der Selbstregulation systematisch eine Auswirkung der Fortbildung auf die drei Basisdimensionen der Unterrichtsqualität zu untersuchen. Zu den großen Stärken unseres Projekts gehören die umfangreiche Schülerstichprobe und das randomisierte, längsschnittliche Kontrollgruppendesign.

Gleichzeitig ergaben sich auch für „Lernen mit Plan“ einige organisatorische und methodische Herausforderungen. So basierte beispielsweise die Anmeldung der Lehrkräfte auf Freiwilligkeit, weshalb die Repräsentativität der Stichprobe und somit die Generalisierbarkeit der Ergebnisse gewissen Einschränkungen unterliegt. Des Weiteren muss in Betracht gezogen werden, dass Angaben von Lehrkräften systematischen Verzerrungseffekten unterliegen können. So könnten Selbstüberschätzung des eigenen

unterrichtlichen Handelns und soziale Erwünschtheit bei den Angaben der Lehrkräfte zu einer positiven Verzerrung der berichteten Veränderungen geführt haben (vgl. Burstein et al., 1995; Clausen, 2002; Porter, 2002). Dies erscheint jedoch eher unwahrscheinlich, da bei den Lehrkräften dann auch eine Veränderung auf den Skalen zur allgemeinen Unterrichtsqualität im Sinne einer *non-equivalent dependent variable* (Shadish, Cook & Campbell, 2002) zu erwarten gewesen wäre. Es kann also davon ausgegangen werden, dass die gefundenen Effekte für die Selbstregulationsförderung im Unterricht nicht auf einen Halo-Effekt (Thorndike, 1920) zurückzuführen sind.

Auch in Bezug auf die Auskünfte der Schülerinnen und Schüler kann der Einfluss von Befragungseffekten nicht ausgeschlossen werden. So wäre es in Anlehnung an Forschungsbefunde zu Panel-Conditioning-Effekten im Rahmen der cognitive stimulus-Hypothese (Sturgis, Allum & Brunton-Smith, 2009) denkbar, dass durch die wiederholte Befragung auf Schülerseite der Unterricht genauer beobachtet sowie kritischer reflektiert und in Folge dessen strenger beurteilt wird. Somit könnten sich zwei Effekte (Interventionseffekt und „Befragungseffekt“) überlagert haben, wobei sich die fortbildungsbedingte Verbesserung der Unterrichtsqualität in einem geringeren Absinken der durch die Experimentalgruppe eingeschätzten Unterrichtsqualität ausgewirkt haben könnte. Solche Effekte ließen sich prinzipiell durch den Einbezug einer dritten Perspektive in Form von Unterrichtsbeobachtung durch entsprechend geschulte Rater zumindest teilweise kontrollieren. Eine zusätzliche (regelmäßige) Unterrichtsbeobachtung ist jedoch in so großen Studien wie „Lernen mit Plan“ kaum realisierbar.

Trotz dieser Einschränkungen stellt der vorliegende Artikel unseres Erachtens einen wichtigen Beitrag zur Aufklärung der Frage dar, inwieweit eine Lehrkräftefortbildung zur Selbstregulationsförderung auch Effekte auf die allgemeine Unterrichtsqualität hat. In zukünftigen Analysen sollte des Weiteren ermittelt werden, inwieweit sich die Fortbildungsteilnahme auch auf die Leistungsentwicklung der Schülerinnen und Schüler ausgewirkt hat und ob die Fortbildungseffekte ausschließlich über die gesteigerte Selbstregulationsförderung oder auch über die Verbesserung der allgemeinen Unterrichtsqualität erklärt werden können.

Literaturverzeichnis

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., et al. (2009). Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente (Materialien aus der Bildungsforschung Nr. 83). Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Gruehn, S., Heyn, S., Köller, O. & Schnabel, K.-U. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU), Dokumentation - Band 1, Skalen Längsschnitt I*. Berlin: Max-Planck-Institut für Bildungsforschung.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*, 463-482.

Borko, H. & Putnam, R. (1995). Expanding a teachers' knowledge base: A cognitive psychologica perspective on professional development. In T. Guskey & M. Huberman (Hrsg.), *Professional development in education: New paradigms and practices* (35-66). New York: Teachers College Press.

Boshuizen, H. P. A. (2004). Does practice make perfect? A slow and discontinuous process. In H. P. A. Boshuizen, R. Bromme & H. Gruber (Hrsg.), *Professional learning: Gaps and transitions on the way from novice to expert* (3-8). Dordrecht: Kluwer Academic Press.

Bromme, R. (1992). Der Lehrer als Experte. Zur Psychologie des professionellen Wissens. Bern: Huber.

Brown, A. L., Campione, J. C. & Day, J. D. (1981). Learning to learn: on training students to learn from texts. *Educational Researcher, 10, 2,* 14–21.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Clausen, M. (2002). Qualität von Unterricht – Eine Frage der Perspektive? Waxmann: Münster.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N. & Orphanos, S. (2009). Professional learning in the learning profession: A status report on teacher development in the United States and abroad. National Staff Development Council.

Deci, E. L. & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum Press.

De Corte, E., Verschaffel, L. & Van de Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. *British Journal of Educational Psychology, 71*, 531–559.

De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction, 10*, 249–266.

Dignath, C. (2009). Different aspects of the promotion of self-regulated learning: a multi-method investigation on the instruction of self-regulated learning at primary and secondary school. Unveröffentlichte Dissertation, Universität Frankfurt.

Dignath, C. & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary level. *Metacognition and Learning, 3*, 231-264.

Fuchs, D. & Fuchs, L. S. (2001). Peer-assisted learning strategies in reading: Extensions for kindergarten, first grade, and high school. *Remedial and Special Education, 22*, 15–21.

Glaser, C., Kessler, C. & Brunstein, J. C. (2009). Förderung selbstregulierten Schreibens bei Viertklässlern: Effekte auf strategiebezogene, holistische und subjektive Maße der Schreibkompetenz. *Zeitschrift für Pädagogische Psychologie, 23*, 5-18.

Gruehn, S. (1995). Vereinbarkeit kognitiver und nichtkognitiver Ziele im Unterricht. *Zeitschrift für Pädagogik, 41*, 531-554.

Hager, W. & Hasselhorn, M. (2000). Psychologische Interventionsmaßnahmen: Was sollen sie bewirken können? In W. Hager, J.-L. Patry & H. Brezing (Hrsg.), *Handbuch Evaluation psychologischer Interventionsmaßnahmen* (41–85). Bern: Huber.

Hattie, John (2008). Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. NY: Routledge.

Hattie, J. A., Biggs, J. & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66*, 99-136.

Helmke, A. (2004): *Unterrichtsqualität: Erfassen, Bewerten, Verbessern* (3. Aufl.). Seelze: Klett-Kallmeyer.

Ihringer, A. (in Vorbereitung). *Transfer-Coaching: Development and Measurement.* Manuskript in Vorbereitung.

Kistner, S., Rakoczy, K.; Otto, B.; Dignath-van Ewijk, C.; Büttner, G. & Klieme, E. (2010). Promotion of self-regulated learning in classrooms: investigating frequency, quality, and consequences for student performance. *Metacognition and Learning, 5*, 157-171.

Klieme, E., Lipowsky, F., Rakoczy, K. & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (127-146). Münster: Waxmann.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung im internationalen Vergleich. In E. Klieme & J. Baumert (Hrsg.), *TIMSS - Impulse für Schule und Unterricht* (43-57). Bonn: Bundesministerium für Bildung und Forschung.

Kline, F. M., Deshler, D. D. & Schumaker, J. B. (1992). Implementing learning strategy instruction in class settings: A research perspective. In M. Pressley, K. R. Harris & J. T. Guthrie (Hrsg.), *Promoting academic competence and literacy in school* (361–406). New York: Academic.

Komorek, E., Bruder, R., Collet, C. & Schmitz, B. (2006). Inhalte und Ergebnisse einer Intervention im Mathematikunterricht der Sekundarstufe I mit einem Unterrichtskonzept zur Förderung mathematischen Problemlösens und von Selbstregulationskompetenzen. In: M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (240-267). Münster: Waxmann.

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (85-113). Münster: Waxmann.

Kunter, M., Brunner, M.; Baumert, J.; Klusmann, U.; Krauss, S.; Blum, W., et al. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. Schulformunterschiede in der Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft, 8,* 502-520.

Labuhn, A. S., Zimmerman, B. J. & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: the influence of feedback and self-evaluative standards. *Metacognition Learning*, *5*, 173-194.

Landmann, M. & Schmitz, B. (Hrsg.) (2007). Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen. Stuttgart: Kohlhammer.

Lipowsky, F. (2010). Lernen im Beruf - Empirische Befunde zur Wirksamkeit von Lehrkräftefortbildung. In F. H. Müller, A. Eichenberger, M. Lüders & J. Mayr (Hrsg.),

Lehrerinnen und Lehrer lernen - Konzepte und Befunde zur Lehrkräftefortbildung (51-72). Münster: Waxmann.

Marks, M., Pressley, M., Coley, J. D., Craig, S., Gardner, R., De-Pinto, W., et al. (1993). Three teachers' adaptions of reciprocal teaching in comparison to traditional reciprocal teaching. *The Elementary School Journal, 94*, 267–283.

Mayer, R. E. (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The Case for Guided Methods of Instruction. *American Psychologist, 59*, 14-19.

Mokhlesgerami, J., Souvignier, E., Rühl, K. & Gold, A. (2007). Naher und weiter Transfer eines Unterrichtsprogramms zur Förderung der Lesekompetenz in der Sekundarstufe I. *Zeitschrift für Pädagogische Psychologie, 21*, 169-180.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6. Auflage). Los Angeles, CA: Muthén & Muthén.

Ophardt, D. & Thiel, F. (2008). Klassenmanagement als Basisdimension der Unterrichtsqualität. In M. K. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion. Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (2. vollst. überarbeitete Auflage, 259-282). Wiesbaden: VS Verlag für Sozialwissenschaften.

Otto, B. (2007). SELVES. Schüler-, Eltern- und Lehrertrainings zur Vermittlung effektiver Selbstregulation. Unveröffentlichte Dissertation, Technische Universität Darmstadt.

Pekrun, R., Götz, T., Jullien, S., Zirngibl, A., v. Hofe, R. & Blum, W. (2002). *Skalenhandbuch PALMA: 1. Messzeitpunkt (5. Klassenstufe).* Universität München: Institut Pädagogische Psychologie.

Perels, F., Gürtler, T. & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction, 15*, 123-139.

Perels, F., Dignath, C. & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education, 24*, 17-32.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*, 3–14.

Rakoczy, K., Buff, A. & Lipowsky, F. (2005). Befragungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 1). Frankfurt a.M.: GFPF/DIPF.

Raghunathan, T. E., Solenberger, P. W. & Van Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software.* Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center.

Reinmann, G. & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (5. vollständig überarbeitete Auflage, 413-458). Weinheim: Beltz PVU.

Rosenshine, B. & Furst, N. (1973). The use 01' direct observation to study teaching. In R. M. W. Travers (Hrsg.), *Second Handbook of Research on Teaching* (122-183). Chicago: Rand McNally.

SAS Institute Inc. (2004). *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Insititute Inc.

Schmitz, B. & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*, 64-96.

Schochet, P. Z. (2008). Guidelines for multiple testing in impact evaluations of educational interventions. Princeton, NJ: Mathematica Policy Research.

Schreblowski, S. & Hasselhorn, M. (2001). Zur Wirkung zusätzlicher Motivänderungskomponenten bei einem metakognitiven Textverarbeitungstraining. *Zeitschrift für Pädagogische Psychologie, 15*, 145–154.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the last decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77, 454-499.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Šidàk, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626–633.

Souvignier, E. & Trenk-Hinterberger, I. (2007). Ein Drei-Stufen-Modell zur Implementation neuer Unterrichtskonzepte in den Schulalltag. In M. Krämer, S. Preiser & K. Brusdeylins (Hrsg.), *Psychologiedidaktik und Evaluation VI* (197–206). Göttingen: V&R unipress.

Spörer, N. & Glaser, C. (2010). Förderung selbstregulierten Lernens im schulischen Kontext. Editorial zum Themenheft. *Zeitschrift für Pädagogische Psychologie, 24*, 171-175.

Sturgis, P., Allum, N. & Brunton-Smith, I. (2009). Attitudes over time: the psychology of panel conditioning. In P. Lynn (Hrsg.), *Methodology of Longitudinal Surveys* (113-126). Chichester, UK: Wiley.

Thorndike, E. L. (1920). A constant error on psychological ratings. *Journal of Applied Psychology*, *4*, 25-29.

Trautwein, U. & Köller, O. (2003). Was lange währt, wird nicht immer gut: Zur Rolle selbstregulativer Strategien bei der Hausaufgabenerledigung. *Zeitschrift für Pädagogische Psychologie, 17,* 199-209.

Vohs, K. D. & Baumeister, R. F. (Hrsg.) (2011). *Handbook of self-regulation: Research, theory, and applications* (2. Auflage). New York, London: The Guilford Press.

Wagner, W., Helmke, A. & Rösner, E. (2009). Deutsch Englisch Schülerleistungen International. Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte (Materialien zur Bildungsforschung. Bd. 25/1). Frankfurt a. M.: GFPF, DIPF.

Wagner, W. (2008). Methodenprobleme bei der Analyse der Unterrichtswahrnehmung und -wirksamkeit - am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz. Dissertation. Universität Koblenz-Landau, Campus Landau, Fachbereich Psychologie.

Waldis, M.; Grob, U.; Pauli, C. & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität - Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (171-208). Münster: Waxmann.

Westfall, P. H. & Young, S. S. (1993). On adusting p-values for multiplicity. *Biometrics, 49*, 941-945.

Wirth, J. (2009). Promoting self-regulated learning through prompts. *Zeitschrift für Pädagogische Psychologie, 23*, 91-94.

Zimmerman, B. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*, 166–183.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Hrsg.), *Self-regulated learning and academic achievement: Theoretical perspectives* (1-37). Mahwah, NJ: Erlbaum.

4

Opening the Black Box: Are Effects of a Teacher Training to Foster Students' Self-Regulation on Students' Math Competencies Mediated by Trained or Untrained Aspects of Teaching Practice?

Werth, S., Wagner, W., Trautwein, U., Lüdtke, O. & Schmitz, B. (zur Einreichung vorgesehene Version des Manuskripts). Opening the Black Box: Are Effects of a Teacher Training to Foster Students-Self-Regulation on Students' Math Competencies Mediated by Trained or Untrained Aspects of Teaching Practice?

Abstract

Teacher trainings that enhance teachers' competence to foster their students' self-regulation during mathematics education are an effective means not only to foster students' self-regulation but have also proved to have positive effects on students' math skills. Still, the mechanisms underlying the effects of such teacher trainings on students' math skills have not yet been investigated. As indicated by current research such teacher trainings might not exclusively affect focal training contents but also general instructional quality comprising e.g. classroom management and teacher support. Thus, the present study investigates whether the effects of a self-regulation enhancing teacher training on students' math skills were mediated by the implementation of focal training contents or by influences on other domains of instructional practice.

The study is based on a sample of 74 classes from the lowest school track in Germany (686 5th graders) who were randomly assigned to either the teacher training or a control condition. Aspects of instructional practice and students' math skills were assessed by student questionnaires and standardized math tests in a pre-post-follow up design. The two-level mediation analyses are based on measures from the pre- and the follow-up test and are conducted in Mplus.

Findings from this study indicate that the training first, affected both focal training contents and untrained aspects of instructional practice and second, had a positive effect on students' math skills. Finally, analyses showed that effects on students' math skills were mediated by a change in focal training contents but not by the change in untrained aspects of instructional practice.

Keywords: teacher training, self-regulation, mediation, math skills

Opening the Black Box: Are Effects of a Teacher Training to Foster Students-Self-Regulation on Students' Math Competencies Mediated by Trained or Untrained Aspects of Teaching Practice?

Self-regulation is an important goal of education and at the same time an essential prerequisite of students' learning and performance (Vohs & Baumeister, 2011; Zimmerman, 2001). Empirical findings indicate that teacher trainings are an effective means to indirectly foster students' self-regulation skills which can ultimately lead to increased achievement (Dignath & Büttner, 2008; Hattie, Biggs & Purdie, 1996). The conclusion drawn from these findings is that the effects of the teacher trainings on students' performance are actually mediated by the implementation of the respective training contents. This conclusion, however, lacks—to our knowledge—empirical confirmation. Most studies that investigate the effects of self-regulation fostering intervention studies separately evaluate the effects of the respective teacher trainings first, on the implementation of training contents and second, on students' performance (Perels, Gürtler & Schmitz, 2005). Further analyses concerning the mechanisms that underlie those observed effects or how these effects are related are unfortunately rarely conducted and to our knowledge until now limited to the area of reading competencies and restricted to analyses on individual level (e.g. Brunstein & Glaser, 2011; Schünemann, Spörer & Brunstein, 2013). This raises the question whether the improvement of math achievement is indeed caused by the successful implementation of the training components that explicitly dealt with self-regulation—or if the achievement gain is based on a more general training effect on aspects such as classroom management.

The current study is based on data from a self-regulation intervention study in Germany that were gathered in a pre-post-follow-up-design. In this study, teachers of 74 fifth-grade classrooms from the lowest school-track ("Hauptschule") were randomly assigned to a training and a control condition. Teachers who participated in the intervention were trained to foster their students' self-regulation during mathematics lessons. Based on student ratings of trained and untrained aspects of instructional quality and on a standardized math test we examine whether the training had an influence on students' math performance and whether these effects were mediated by a change in focal training contents (promotion of self-regulation-enhancing aspects of instructional quality) or a change in untrained aspects of instructional quality (classroom organization, instructional support, emotional support).

Effects of teacher trainings to foster students' self-regulation

Self-regulation or self-regulated learning can be defined as "an active, constructive process, whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment" (Pintrich, 2000, p. 453). Empirical findings show that self-regulated learning can be fostered by teacher trainings that combine implicit and explicit promotion of self-regulation (Dignath & Büttner, 2008). The explicit promotion of students' self-regulation implies that teachers introduce self-regulation strategies to their students and train these strategies with their students (Dignath, 2009). The implicit promotion of self-regulation strategies, in contrast, implies that teachers shape a self-regulation enhancing learning environment including not only self-regulation enhancing tasks and teaching methods but also prompts that stimulate students' application of recently learned self-regulation strategies (see Dignath, 2009). A combination of both explicit and implicit promotion of students' self-regulation is assumed to be most effective with regard to a sustainable acquisition of self-regulation strategies by students and an improvement of their performance (Kistner et al., 2010).

Furthermore, empirical findings indicate that a combination of self-regulation strategies with mathematical content is especially effective not only with regard to promoting students' self-regulation but also to enhance students' math competencies (Dignath & Büttner, 2008; Hattie, 2008). Effects of such self-regulation enhancing teacher trainings on students' math competencies showed effect sizes of $ES \geq 0.96$ (Dignath & Büttner, 2008) depending e.g. on the age-group of the students, the underlying theoretical model and the duration, intensity, and quality of the teacher training (Dignath & Büttner, 2008). Studies that investigate the effects of teacher trainings that fosters the self-regulation enhancing quality of instruction on students' math performance implicitly assume the focal training contents to mediate the impact of the training on student outcomes (De Corte, 2000; Perels, Dignath & Schmitz, 2009). Still, this assumed mediation is usually not tested (Spörer & Glaser, 2010), i.e. it has not yet been investigated whether the increase of students' performance after such teacher trainings was indeed linked to the implementation of focal training contents (e.g. the promotion of meta-cognitive strategies) in daily teaching practice (e.g. Souvignier & Trenk-Hinterberger, 2010). Consequently, it has neither been investigated whether potential effects on students math performance were a consequence of side-effects on other aspects of instructional quality which may occur when the contents of self-regulation enhancing

programs have to be integrated in already existing patterns of instructional quality (Werth et al., 2012).

Effects of teacher trainings to foster students' self-regulation on general aspects of instructional quality

Instructional quality which has been defined as comprising all classroom interactions between teachers and students that stimulate students' learning and motivational development (Caroll, 1963; Perels, Gürtler & Schmitz, 2005; Weinert, Schrader & Helmke, 1989) is a crucial factor for students‘ achievement. According to Pianta, La Paro, and Hamre (2008; see also Baumert et al., 2004; Klieme, Pauli & Reusser, 2009) instructional quality can be described by three domains: class organization, instructional support, and emotional support. According to this classification, class organization comprises aspects such as classroom management and time on task (Klieme et al., 2009). Instructional support covers teaching components such as discursive practice and activating prior knowledge and relates to any instructional practice that fosters students' higher-level thinking (Klieme et al., 2009; Seidel, Rimmele & Prenzel, 2005). Emotional support covers all emotional/affective aspects of teacher-learner interactions such as support of autonomy, participation, and teacher support (Klieme et al., 2009). Especially the first and the second domain have been shown to directly affect students' cognitive growth and learning processes (Klieme, Lipowsky, Rakoczy & Ratzka, 2006; Klieme et al., 2009). In contrast, emotional support has an indirect effect on students' achievement by affecting motivational variables (Klieme et al., 2006; Klieme et al., 2009; Lipowsky et al., 2009).

Empirical findings indicate that teachers vary concerning their pattern of instructional quality and that aspects of instructional quality can be changed by teacher trainings (Borko & Putnam, 1995; Desimone, Porter, Garet, Yoon & Birman, 2002; Wei, Darling-Hammont, Andree, Richardson & Orphanos, 2009; Yoon, Duncan, Lee, Scarloss & Shapley, 2007). However, the influence of teacher trainings on instructional practice and student outcomes is not yet fully explored, with several studies showing and discussing effects that were smaller than expected (e.g. Garet et al., 2011). As outlined by researchers like Berliner (2001), Boshuizen (2004), Borko und Putnam (1995); Bromme (1992) the integration of new teaching practices might require teachers to modify former teaching style and disturb established teaching routines. These considerations let us, on the one hand, assume that a lack of time and capacity on side of the teachers might lead to unintended and unexpected negative changes in prior teaching practice such as a decrease of discipline in the classroom while applying

recently learned teaching methods . On the other hand, the participation in a well-designed teacher training can also support teachers' classroom organization by providing them with well-designed teaching materials and scripted lessons (De Corte, 2000; de Jager, Reezigt & Creemers, 2002; Fuchs et al., 2003; Perels, Dignath & Schmitz, 2009). Such a positive effect has been shown in a study by Werth et al. (2012) where a teacher training which aimed at fostering the self-regulation enhancing learning environment not only had positive effects on training contents but also affected untrained instructional features that belong to the aforementioned domains classroom organization, instructional support, and emotional support. Considerably different effect sizes for trained and untrained aspects in this study indicate that there is no reason to suspect that the reported improvement of the assessed domains of instructional quality was an artifact caused by a global rating tendency or a global perception on side of the students. Thus, a potential explanation given by the authors was that the supply of scripted lessons might have supported teachers in making their lessons more effective.

Having in mind the effects of the three domains of instructional practice on students' cognitive and motivational development such potential side-effects on untrained aspects of instructional practice might have an additional unintended effect on students' learning. This has the consequence that potential effects of the intervention on student achievement can't clearly be linked to the implementation of program contents (Rossi, Lipsey & Freeman, 2004; Yoon, Duncan, Lee, Scarloss & Shapley, 2007). However, the question whether the reported effects of such trainings on students' math achievement are actually mediated by the implementation of focal training contents or by a change of untrained aspects of instructional quality has not yet been investigated empirically.

Hypotheses

In line with prior research the study investigates the effects of a teacher training that aimed at enhancement of teachers' ability to foster their students' self-regulation has effects on students' math achievement. The present study focuses on the actual mechanisms that underlie the effects of such a training on students' math achievement, i.e. whether the effects are indeed mediated by the implementation of self-regulation fostering aspects of instructional quality (abbreviated *siq)* or whether the effects are mediated by effects of the training on other aspects of general instructional quality (abbreviated *giq*) that were also affected by the training. The underlying intervention comprised a four-day teacher training that provided

teachers with teaching materials and scripted lessons. The effects of the program were assessed by a pre-post-follow-up measurement.

In line with research about the effects of self-regulation fostering teacher trainings on students' math competencies (Dignath & Büttner, 2008) we expected a positive effect of the teacher training on students' math performance in the follow-up test (t3) (hypothesis 1). Based on prior research (e.g. Dignath & Büttner, 2008), we also assumed that the participation in the teacher training would have a positive impact on the implementation of self-regulation enhancing aspects of instructional quality at t3 (hypothesis 2). Furthermore, we assumed that—as previously shown by Werth et al. (2012)—the teacher training in the current study would positively affect untrained aspects of instructional quality at t3 (hypothesis 3). As self-regulation enhancing aspects of instructional quality are supposed to have a positive influence on students' math performance (e.g. Hattie, Biggs & Purdie, 1996), we assumed that the expected positive effect of the training on students' math competencies was mediated by the implementation of the focal training contents (hypothesis 4). However, as also the other, untrained domains of general instructional quality are supposed to have a positive influence on students' math (Lipowsky et al., 2009) performance, we additionally assumed that the effect of the teacher training on math performance was also mediated by the untrained aspects of instructional quality (hypothesis 5). The assumed mediation model is depicted in Figure 1

Method

Procedure

The present analyses are based on data from a randomized experimental study about the effects of a teacher training program on students' self-regulated learning in math lessons (cf. Werth et al., 2012). The teacher training program in total comprised 13 weeks of schooling and included a four-days teacher training that was delivered in two blocks, one at the beginning of the intervention and one block four weeks after the first training session. Between the first and the second training session and between the second training session and the end of the intervention the teachers had to implement the focal training contents in their regular lessons. The selection of focal training contents was based on the self-regulation model by Schmitz und Wiese (2006) and covered seven aspects of the self-regulation process (goal clarity, planning, self-motivation, volition, handling errors and frame of reference, and reflection).

The implementation of those training contents was supported by a teacher manual that contained one scripted lesson per week in order to explicitly foster students' self-regulation without, however, relating these strategies to mathematics. Additionally the manual contained information on how to implicitly foster students' self-regulation by making prompts (Thillmann, Künsting, Wirth & Leutner, 2009) and asking students to apply the already trained strategies to mathematics education. For instance, the suggestion for implicitly promoting students' goal setting was to mention the goals for the lesson or the week and to request the students in the beginning of each school week to set own, individual goals for the respective week (e.g. "I want to put up my hand at least five times this week."".)

Data used in this study were based on student questionnaires and a standardized math test that were administered three times across the 13 weeks of schooling between the pretest and the follow-up measurement. The first measurement (t1) took place in February and the third measurement (t3) occurred in June.

Sample

The present student sample consisted of 681 students in 74 classes from the lowest school track ("Hauptschule") in Germany. Due to time constraints, only about half of the students in each class were asked to complete questions about instructional quality, whereas the other students worked on items that are not relevant in the present context. Thus, the average number of students per class was 9.11 students. The average age was 10.95 years (SD = 0.80), and gender was distributed almost equally (51.7 % boys).

Instruments

The student questionnaires included several scales that asked students to rate their mathematics lessons/teacher. At each measurement time point students were asked to keep in mind the previous 4 weeks, which approximately corresponds to the time period between the measurement time points of the applied pre-post-follow-up-design. For the purpose of the present investigation, we focused on a subset of five scales gathered at t1 and t3. Whereas two scales (*goal clarity*, *handling errors*) measured the implementation of focal training contents, the other three scales (*classroom management*, *discursive practice, teacher support*) measured aspects of instructional quality that were not explicitly targeted by the training, each scale representing one of the three domains of instructional quality postulated by Baumert et al. (2004), (Klieme, Pauli & Reusser, 2009), and (Pianta, La Paro & Hamre, 2008). Further information concerning the applied scales and items can be gathered from Table 1.

Table 1

Descriptives, Reliabilities, Intraclass Correlations, and Number of Items for the Applied Scales

Construct		t1			t3			Number of items
		M(SD)	Cronbach's α	ICC	*M(SD)*	Cronbach's α	ICC	
Self-regulation enhancing aspects of instructional quality (focal training contents)								
Goal clarity	*CG*	3.08 (0.20)	.75	0.10	2.72 (0.23)	.79	0.08	5
	EG	2.92 (0.20)			2.89 (0.17)			
Handling errors	*CG*	3.14 (0.19)	.71	0.09	2.68 (0.30)	.81	0.14	4
	EG	3.07 (0.21)			2.88 (0.30)			
Untrained aspects of instructional quality								
Classroom management	*CG*	2.79 (0.36)	.76	0.17	2.65 (0.37)	.80	0.11	4
	EG	2.65 (0.22)			2.70 (0.20)			
Discursive practice	*CG*	3.06 (0.17)	.73	0.08	2.75 (0.27)	.81	0.08	6
	EG	3.05 (0.15)			2.92 (0.15)			
Teacher support	*CG*	3.42 (0.23)	.79	0.10	2.90 (0.31)	.80	0.08	4
	EG	3.25 (0.16)			2.99 (0.15)			

Note. Sample: CG = Control group (n = 258–263), EG = Experimental group (n = 427–432). Response format: 1 = *I don't agree at all*; 2 = *I don't agree*; 3 = *I agree*; 4 = *I completely agree.*

Scales for measuring the fostering of self-regulation during math lessons

The scale goal clarity describes the extent to which the teacher clearly mentions his/her goals for the lesson, the week, or the year. As the students were explicitly told to set own goals the teacher' goal clarity can be counted as a means to foster students' self-regulation. In the present investigation, goal clarity was assessed with five items taken from the TRAIN study inventory (Jonkmann, Rose & Trautwein, 2013) and the PYTHAGORAS study (Rakoczy, Buff & Lipowsky, 2005) (sample item: "At the beginning of the lesson, our math teacher clearly explains what s/he wants us to learn."). The internal consistency (Cronbach's alpha) of the scale was $\alpha \leq .75$. About 10% of the variability was at the class level (ICC_{t1} = .10; ICC_{t3} = .08).

Handling errors is a scale that captures the extent to which the teacher deals with students' errors in a constructive way and clearly points to the possibility to learn from mistakes. Since the topic of how to constructively handle errors was one aspect of the teacher training it was counted as focal training content. In the present investigation, handling errors was assessed with four items taken from the DESI study inventory Wagner, Helmke, and Rösner (2009; sample item: "My teacher says that errors are useful because we can learn something from them."). The internal consistency (Cronbach's alpha) of the scale was $\alpha \leq .71$. About 10% of the variability was at the class level (ICC_{t1} = .09; ICC_{t3} = .14).

Scales for measuring general instructional quality during math lessons

The construct classroom management taps the occurrence of disciplinary problems and belongs to the aforementioned domain classroom organization. In the present investigation, classroom management was assessed with four items that were adapted from the PISA 2003 (Ramm et al., 2006) and COACTIV studies Baumert et al. (2009; sample item: "There are usually no disturbances during our math teacher's lessons"). The internal consistency (Cronbach's alpha) of the classroom management scale was $\alpha \leq .76$. More than 10% of the variability was at the class level (ICC_{t1} = .17; ICC_{t3} = .11).

The scale discursive practice taps how much the teacher fosters his/her students' active thinking and belongs to the aforementioned domain cognitive activation. In the present investigation, discursive practice was assessed with six items that were developed by Rakoczy et al. (2005; sample item: "Our teacher wants us to find different ways for solving a task."). The internal consistency (Cronbach's alpha) of the classroom management scale was $\alpha \leq .73$. Relative agreement was acceptable (ICC_{t1} = .08; ICC_{t3} = .08).

The scale teacher support taps how much the teacher supports his/her students during the learning process and belongs to the aforementioned domain emotional support. In the present investigation, teacher support was assessed with four items that were adapted from the TRAIN study inventory (Jonkmann, Rose & Trautwein, 2013) which, in turn, refers to the PISA 2003 (Ramm et al., 2006) and COACTIV studies (Baumert et al., 2009) (sample item: "Our teacher encourages us to ask when we don't understand something."). The internal consistency (Cronbach's alpha) of the teacher support scale was $\alpha \leq .79$. Relative agreement was acceptable ($ICC_{t1} = .10$; $ICC_{t3} = .08$).

In order to assess students' learning outcomes we conducted a standardized math test at the three aforementioned time points t1, t2, and t3. The test took 35 minutes and comprised 34 items with mostly closed response format. In order to prevent cheating the test was applied in two parallel test forms which only differed concerning the order of the tasks. Both test forms were equally distributed in each class. In order to make pre-, post- and follow-up scores comparable each measurement time point contained two blocks of 10 anchor items. The content of the test was in line with the respective curriculum. Item- and person- parameters for students' math competence were estimated with a 2-PL model using Mplus 6.0 (Muthén & Muthén, 1998-2010). The complete model was tested by estimating a three-dimensional IRT model (one dimension for each measurement time point including measurement invariance across measurement occasions) using Bayesian parameter estimation. The posterior predictive *p*-value was $p = 0.102$, indicating good model-fit. Expected *a posteriori* person-parameter estimates (EAPs) were used for further analyses. About 15% to 18% of the variability was at the class level ($ICC_{t1} = .18$; $ICC_{t3} = .15$).

Statistical Analyses

As outlined before, our study examined the effect of a teacher training to foster students self-regulation on math competencies and whether this effect was exclusively mediated by the implementation of focal training contents or also by aspects of instructional quality that were also affected by the intervention.

According to the causal steps approach by Baron and Kenny (1986; Judd & Kenny, 1981; MacKinnon, 2008) a mediation effect must be tested against four conditions (see Figure 1). First, it hast to be shown, that the independent variable (in our case the participation in the teacher training: *treatment*) actually has an effect *c* on the dependent outcome variable (in our case the students' math competence: *math*). Second, it has to be shown that the treatment also has an effect *a* on the potential mediators (in case of our study the composite for focal training

contents *siq* or the composite for untrained aspects of general instructional quality *giq*). In the full mediation model, it has finally to be shown that these potential mediators also have an effect *b* on the dependent outcome variable and that the direct effect *c'* (formerly *c*) of the treatment on the outcome variable disappears or is reduced under control of the mediator while path *a* and *b* are consistent and significant. Although MacKinnon, Lockwood, Hoffman, West und Sheets (2002) showed that actually only path *a* and path *b* are required to be significant in order to assume an indirect effect we will test all paths that are included in the full mediation model because the interpretation of the mediated effect is clearer if the full mediation model can be confirmed (MacKinnon, 2008). The size of the indirect effect *ab* can be estimated and tested by applying the model indirect command which is implemented in Mplus (Muthén & Muthén, 1998-2010). A further condition for the assumption of a mediation effect is the correct temporal order of the constructs that were assessed, i.e. the independent variable captures something that happened first, the mediator captures something that happened thereafter but before the construct that is captured by the dependent variable (MacKinnon, 2008). In our case the independent variable, i.e. the teacher training took place before t3. The mediator, i.e. the implementation of focal training contents and the change of untrained aspects of instructional quality were assessed by asking the students at t3 about the previous four weeks of instruction and thus captures something that happened before t3. Math competencies were measured at t3 and reflect the actual competence of the students at the day of the test.

Statistical testing was conducted in Mplus 6 which only allows for two-tailed testing (Muthén & Muthén, 1998-2010). In accordance to conventional guidelines for statistical testing, the acquainted *p*-values were then interpreted along the criteria for one-tailed testing (i.e. a 10 percent alpha level) because all hypotheses assumed directed effects. To prevent alpha-inflation when testing several potential mediators, we aggregated the respective scales—based on the assumptions concerning the effects of the training on these scales—to form two composites (Schochet, 2008), one for the focal training contents *siq* and one for the untrained aspects of general instructional quality *giq*. In order to prevent the indirect effects from cancelling each other out by having opposite signs (Hayes, 2009) and in order to prevent biased regression coefficients due to multicollinearity each of these two composites was then tested in separate models against the four criteria of a full mediation model outlined before.

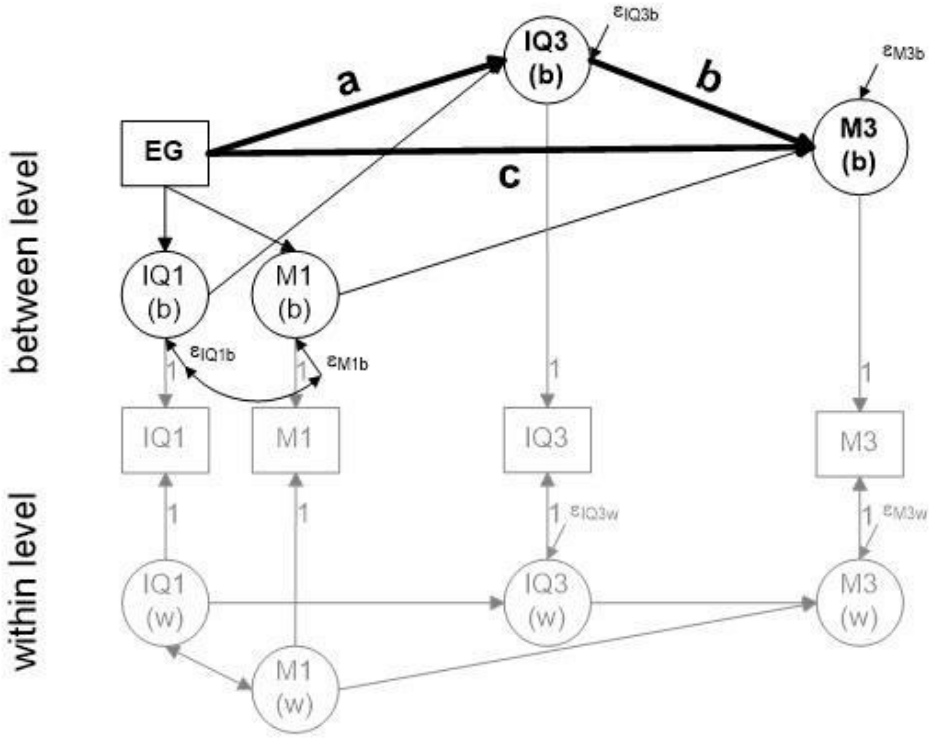

Figure 1. Two-level full mediation model correcting standard errors for clustering at the class level. Estimates are from Mplus 6.0 (Muthén & Muthén, 1998–2010), computed with the model indirect command: within = individual student ratings; between = student ratings at the class level. The letters EG stand for experimental group and thus for one of the two categorie of the treatment (treatm); the letters IQ stand for instructional quality and comprise focal training contents (sig) and general instructional quality (giq). Fit indices for the two separate models:

Model 2a (s*iq t3* as mediator): χ^2 (df = 4, N = 681) = 8.07, CFI = .99, RMSEA = .039, $SRMR_{within}$ = .010; $SRMR_{between}$ = .027

Model 2b (*giq t3* as mediator): χ^2 (df = 4, N = 681) = 2.81, CFI = 1.00, RMSEA = .000, $SRMR_{within}$ = .017; $SRMR_{between}$ = .019.

As the separate estimation of the indirect effects might also lead to alpha-inflation due to multiple testing, we afterwards adjusted the respective *p*-values of the indirect effects by using the Šidák-correction (Šidák, 1967) $p_{adjusted} = 1 - (1 - p_{unadjusted})^k$, while k corresponds to the number of tests. In order to control for pre-test scores we applied an ANCOVA approach

with the t1 measures as covariates (MacKinnon, 2008)[15]. To address missing data (the coverage rate for questionnaire items and test scores ranged from 81% to 87%) we used the full information maximum likelihood estimator (FIML; Enders & Bandolos, 2001; Muthén & Muthén, 1998-2010). All models were set up as multi-level models using the twolevel-command in Mplus to account for the hierarchical structure of the data, i.e. that students are nested in classes. Furthermore, multi-level modeling helps to separate variance on the individual level from variance on the class-level. Using a notation convention suggested by Krull und MacKinnon (2001) our models can be categorized as 2-1-1 mediation models that are restricted to the class level (Pituch & Stapleton, 2012). This notation implies that the independent variable (treatment condition) doesn't vary within classes and is thus notated as a variable on level 2 or class-level, while the respective mediator (perceived change of instructional quality due to the implementation of focal trainings contents or unintended side-effects) and the dependent variable (students' math competence) differ between individual students within classes and are thus notated as level-1 variables (Preacher, Zhang & Zyphur, 2011).

Based on the aforementioned considerations we specified three models in order to test our hypotheses. In Model 1, the total effect model, we tested the assumption that the teacher training actually had an effect on students' math performance at t3. In Model 2a, we tested whether path *a* is significant, i.e. whether the treatment had an effect on *siq*-composite at t3. In a separate Model 2b, we tested path *a* for the *giq*-composite. In Model 3a and 3b, which is the full mediation model for each composite, we regressed the dependent variable on both the independent variable and on the respective mediator (i.e. the composites *siq* or *giq*). In these models we tested path *b*, i.e. whether the potential mediators have an effect on the dependent outcome variable and whether the effect of the independent variable on the outcome variable (path *c'*) is smaller than path *c* or even disappears. In this model we further estimated and tested the respective indirect effect *ab* with the Sobel test (first-order delta method) by using the model indirect command in Mplus (Muthén & Muthén, 1998-2010). Finally, the indirect effects were additionally tested with the joint test of significance which tests the null hypothesis ($ab = 0$) by testing that both paths *a* and *b* are zero. According to Kenny an advantage of this test is that it performs as well as the bootstrap test (Hayes & Scharkow, 2013) which is less conservative than the Sobel test but is unfortunately not implemented in

[15] In contrast to the analyses conducted by Werth et al. (2012), we applied an ANCOVA approach with the t1 measures as covariates instead of using a change score approach (MacKinnon, 2008; Maxwell & Delaney, 2004). This procedure was chosen to account for baseline imbalance in some of the pre-test scores that occurred despite randomization (Kisbu-Sakarya, MacKinnon & Aiken, 2013).

Mplus (when using the twolevel command). However, as this test presumes that *a* and *b* are uncorrelated and does not provide a confidence interval for the indirect effect it should only be used in combination with other tests (Fritz, Taylor & MacKinnon, 2012).

Results

As outlined in the previous section, we set up three multi-level models to address our research questions. The results in the following section shall be presented along our five hypotheses:

Hypothesis 1: We expected a positive effect of the teacher training on students' math performance at t3.

Hypothesis 2: We assumed that the participation in the teacher training would have a positive impact on the implementation of self-regulation enhancing aspects of instructional quality at t3.

Hypothesis 3: We assumed that the teacher training in the current study would positively affect untrained aspects of instructional quality at t3.

Hypothesis 4: We assumed that the expected positive effect of the training on students' math competencies was mediated by the implementation of the focal training contents.

Hypothesis 5: We assumed that the effect of the teacher training on math performance was also mediated by the untrained aspects of instructional quality.

Model 1: Effect of the training on math performance

To test hypothesis 1 Model 1 examined the total effect *c* of the independent variable on the educational outcome by regressing math performance at t3 on the dummy coded treatment condition *treatm* (0 = control group; 1 = treatment condition) while controlling for the math score at t1. As depicted in Table 2 there was a significant positive effect of the teacher training on students' math-performance ($\beta c = .44$, $p < .05$).

Model 2a and 2b: Effect of the training on the mediators, effect of the mediators on math performance and full mediation model

To test hypothesis 2, 3, 4, and 5 we set up two full mediation models, one for each construct. Each of these models included the treatment, one of the mediators (*siq t3* or *giq t3*), and the math score at t3 as the outcome variable. Additionally we included the t1 scores of the respective mediator and the outcome variable. In these models the respective mediator was regressed on the treatment condition while controlling for the t1-score of the mediator. As depicted in Table 2, the included covariates were significantly related to the respective t3

scores. Furthermore, Table 2 shows that both mediators were positively affected by the treatment, whereby the effect of the training on the focal training contents (βa_{siq} = 1.34, p < .001) was slightly stronger than the effect of the training on the untrained aspects of instructional quality (βa_{giq} = 1.02, p < .001).

Table 2

Results of the total effect model and the full mediation model

	Total effect model		Full mediation model 2a				Full mediation model 2b			
	Treatment → Math t3		Treatment → SIQ t3		Treatment, SIQ t3 → Math t3		Treatment → GIQ t3		Treatment, GIQ t3 → Math t3	
	βc	*p*	βa_{siq}	*p*	βb_{siq}	*p*	βa_{giq}	*p*	βb_{giq}	*p*
Treatment	0.44	0.042 [a]	1.34	0.000 [a]	0.11	0.338 [a]	1.02	0.001 [a]	0.34	0.092[a]
SIQ/GIQ t3	_	_	_	_	0.47	0.002 [a]	_	_	0.29	0.023[a]
Math t1	0.77	0.000 [a]	_	_	0.74	0.000 [a]	_	_	0.70	0.000[a]
SIQ/GIQ t1	_	_	0.92	0.000 [a]	_	_	0.91	0.000 [a]	_	_
Indirect Effect	_	_	_	_	0.63	0.019 [ab]	_	_	0.30	0.099[ab]

Notes. Sample total effect model: 680. Sample full mediation models: 681. The treatment condition was coded 0 = control group, 1 = experimental group. Regression coefficients β were semi-standardized in case of dummy-coded variables and indirect effects and fully-standardized for metric variables; the p-values were drawn from the unstandardized output.

[a]. *P*-value was divided by two because all *p*-values were gained from two-sided tests but have to be interpreted for one-sided testing.

[b]. *P*-value with adjustment for multiple testing.

To test whether full mediation can be assumed for both mediators, in our models math performance at t3 was separately regressed on each of the mediators *siq t3* and *giq t3*. In both models the mediators were controlled for their t1 scores. Additionally, math performance at t3 was controlled for the math score at t1. Both covariates were themselves regressed on the treatment variable to control for potential differences at t1. Furthermore, in each model the covariates were allowed to be correlated. Model 2a and 2b are depicted in Figure 1. As indicated by the fit-indices below Figure 1 both models proved to have a good fit.

As depicted in Table 2, both full mediation models meet all criteria suggested by Baron and Kenny (1986) to assume the existence of a full mediation. First, math performance at t3 was positively related to *siq t3* (βb_{siq} = .47, $p < .05$) and *giq t3* (βb_{giq} = .29, $p < .05$). Second, the effect *c'* of the treatment on math performance at t3 is non-significant when the respective mediators are included in the model while the other paths *a* and *b* are consistently significant. When separately quantifying the size of the indirect effects of both mediators the indirect effect of the *siq* composite (βab_{siq} = .63, $p < .05$) proved to be slightly stronger than the indirect effect of the *giq* composite (βab_{giq} = .30, $p < .10$). However, after the application of the Šidak-correction (Šidak, 1967) which prevents alpha-inflation by adjusting the alpha-niveau for multiple testing the new alpha level for testing both mediation effects in separate models was $p_{adjusted}$ = 0.025. With regard to this adjusted criteria of significance the effect of the *siq* composite still proved to be significant (p = .019), while the effect of the *giq* composite was no longer significant (p = .099).

Discussion

With regard to these results it can be concluded that the teacher training in the present study indeed positively affected teachers' implementation of focal training contents and students' math achievement. This confirms results from prior research that was e.g. presented in the meta-analysis of Dignath und Büttner (2008) who found self-regulation enhancing teacher trainings to have an effect on student's math skills. Furthermore, the study showed that the effect of the training on student outcomes was mediated by the implementation of focal training contents in daily teaching and is thus in line with the implicit assumption of researchers when setting up a teacher training that aims at fostering students' self-regulation. However, in line with (Werth et al., 2012) the study additionally shows that the participation in the teacher training also affected domains of teachers' general teaching practice like classroom management, discursive practice and teacher support that were not addressed by the teacher training. In line with research that assumes (changes in) general teaching practice

to also influence students' math achievement (Klieme, Pauli & Reusser, 2009) we found an indirect effect of the teacher training on students' math competencies that is delivered by the positive influence of the intervention on general teaching practice. However, due to our decision to carefully control for the fallacies of multiple testing this concurrent indirect effect can’t be confirmed because the p-value wasn’t significant after correcting for the number of conducted tests.

Concerning the mechanisms behind the effect of self-regulation-related teacher trainings on students' math achievement the present study indicates the crucial importance of not only investigating the effects of focal training contents on the respective outcome but to also consider effects on other aspects of teaching quality and their potential influence on the measured outcome. As mentioned by Spörer und Glaser (2010) the successful implementation of focal training contents (e.g. the "degree of fidelity in the implementation of the [...] treatment[...]" in the study of Brunstein and Glaser (2011, p. 928)) often doesn't lead to the expected effects on student outcomes— in case that the implementation of focal training contents is measured at all (Spörer & Glaser, 2010). Furthermore, only few studies (Brunstein & Glaser, 2011; Schünemann, Spörer & Brunstein, 2013) investigated whether it was indeed students' adoption of components of self-regulated learning that affected the outcomes. Being aware that the implementation of less-structured training contents might even have a negative effect on instructional quality (Kline, Deshler & B.Schumaker, 1992), the unexpectedly small effects could be a consequence of a positive effect of such trainings on students' self-regulation and a simultaneous negative effect on instructional quality that level each other out.

As shown by several researchers like Darling-Hammond, Wei, Andree, Richardson und Orphanos (2009) and Desimone, Porter, Garet, Yoon und Birman (2002) the transfer of training contents in regular teaching practice seems to be a highly complex and fragile process that requires e.g. a minimum duration of the training, a combination of theory and practice and the provision of well-structured teaching materials. Being aware of these high requirements of efficacious teacher trainings many self-regulation trainings use trained assistants instead of teachers to deliver the treatment (e.g. Brunstein & Glaser, 2011; Schünemann et al., 2013). This goes along with the question whether comparable effects could be gained by training teachers instead of assistants who then deliver the treatment to their students. Only some self-regulation related studies actually trained teachers to deliver the treatment and only few of them were designed sufficiently intense in order to change teachers' teaching practice or even more distal outcomes like students' math or reading competencies (e.g. De Corte, 2000; Perels, Dignath & Schmitz, 2009). Most of these studies, however, only

realized small sample sizes without a control group and did thus not allow for analyses on class level or the test of mediational hypotheses. In contrast, our study not only trained teachers to deliver the treatment but also realized a pre-post-follow-up randomized controlled trial and can thus link the effects of the intervention to effects on students' outcomes. Our study can thus be assumed to have high ecological validity and practical usability as the effects took place in ordinary classrooms (Schünemann et al., 2013).

Limitations and further research

Despite the exceptionally strong design of our study our analyses were unfortunately faced with several challenges. Some of these limitations were caused by idiosyncratic problems of the underlying study that shall be briefly described. However, there were also some more severe methodological challenges that shall be described in more detail, as they can be regarded as essential challenges to research in this field.

As shortcoming of our study it must be mentioned that our analyses could only rely on mediators and dependent variables from measurement time point three. This, however, can be explained by referring to research from the field of teacher training research that indicated that the effects on teaching practice and students' outcomes have to be expected with some retardation. Even taking the mediator and the dependent variable from the same measurement time point is in case of our study not problematic. Since the prompt of the mediator asked students to report about the teaching practice in the past weeks between the measurement time point 2 and measurement time point 3 the required chronological order of the mediator and the dependent variable could still be met (MacKinnon, 2008).

More severe challenges had to be faced due to the high intercorrelation between student ratings of different constructs on class level. First, the high correlation raises doubts concerning the validity of our measures, since the criterion of discriminant validity (Greenwald, 1997) is not met. Yet, referring to research of Wagner (2008) (see also Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013) we have to assume that also other studies had to face this challenge which Wagner (2008) explained by a biased global perception of the teacher. Furthermore, the different regression weights for both composites, i.e. slightly smaller regression coefficients for the composite general instructional quality indicates that student perception was influenced by a global perception but still recognized differences between both composites. In line with Fauth, Decristan, Rieser, Klieme und Büttner (2014) we, thus, assume that there might be some halo bias in the ratings, but that this does not threaten the validity of our findings. Second, the high intercorrelation of the single constructs leads to multicollinearity and thus to suppression effects when setting up a model that

includes both mediators and their pretest values. In order to prevent our analyses from suppression effects going along with biased standard errors we tested both mediators in separate models and adjusted for multiple testing afterwards.

As a last point, it would have been interesting to also include teacher ratings in our analyses. Although student ratings can be regarded as valid indicators of instructional quality even in case of students from primary school, there is also the finding that ratings from students from lower school tracks are less sensitive to didactical aspects when rating instructional quality than students from higher school tracks. Since the students in our sample were not only quite young and additionally from the lowest school track in Germany, it would have been interesting to also include teacher ratings in the analyses which are known to be more sensitive to didactical aspects. However, due to the relatively small sample size it was not possible to include teacher ratings in our analyses. Furthermore, studies by (Werth et al., 2012) indicated that teachers might perceive the influences of teacher trainings differently than students. Altogether the restriction to student ratings was acceptable in our study as student ratings are known to have a high predictive influence on student achievement (Fauth et al., 2014). However, the inclusion of teacher ratings would help to draw even a clearer picture of the mechanisms that underly teacher trainings that aims at the enhancement of students self-regulation and by thus an improvement of students' math achievement.

Altogether the study can be regarded as an important contribution to depicting and understanding the actual mechanisms behind the effects of self-regulation enhancing teacher trainings on students' math competencies. The implication for practice from this study is that the complexity of teacher trainings as a means of school development must not be underestimated. When designing intervention studies that train teachers in order to deliver a training to their students it is thus not only important investigate the critical components of the intervention and the hypothesized mediating process (cf. Schünemann, Spörer & Brunstein, 2013) but also potential side-effects on regular teaching practice that has itself an influence on students learning.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182. doi: 10.1037/0022-3514.51.6.1173

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., . . . Tsai, Y.-M. (2009). Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente. [Professional competence of teachers, cognitively activating instruction, and the development of students' mathematical literacy (COACTIV): Documentation of the instruments]. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W., & Neubrand, M. (2004). Mathematics teaching from the perspective of the PISA students and their teachers. In The German PISA Consortium (Ed.), *PISA 2003: The educational level of adolescents in Germany—the second international comparison* (pp. 314-354). Münster, Germany Waxmann.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463-482. doi: http://dx.doi.org/10.1016/S0883-0355(02)00004-6

Borko, H., & Putnam, R. T. (1995). Expanding a teachers' knowledge base: A cognitive psychologica perspective on professional development. In T. R. Guskey & M. Huberman (Eds.), *Professional development in education: New paradigms and practices* (pp. 35–66). New York: Teachers College Press.

Boshuizen, H. P. A. (2004). Does practice make perfect? A slow and discontinuous process. In H. P. A. Boshuizen, R. Bromme & H. Gruber (Eds.), *Professional learning: Gaps and transitions on the way from novice to expert* (pp. 3-8). Dordrecht: Kluwer Academic Press.

Bromme, R. (1992). Der Lehrer als Experte : zur Psychologie des professionellen Wissens. Bern {[u.a.]: Huber.

Brunstein, J. C., & Glaser, C. (2011). Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology, 103*(4), 922-938. doi: 10.1037/a0024622

Caroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession. A status report on teacher development in the United States and abroad. Dallas, TX: National Staff Development Council.

De Corte, E. (2000). Marrying theory building and the improvement of school practice: a permanent challenge for instructional psychology. *Learning and Instruction, 10*(3), 249-266. doi: http://dx.doi.org/10.1016/S0959-4752(99)00029-8

de Jager, B., Reezigt, G. J., & Creemers, B. P. M. (2002). The effects of teacher training on new instructional behaviour in reading comprehension. *Teaching and Teacher Education, 18*(7), 831-842. doi: http://dx.doi.org/10.1016/S0742-051X(02)00046-X

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81-112. doi: 10.3102/01623737024002081

Dignath, C. (2009). Different aspects of the promotion of selfregulated learning: a multi-method investigation on the instruction of self-regulated learning at primary and secondary school. Johann Wolfgang Goethe-Universität, Universitätsbibliothek Frankfurt am Main. Retrieved from urn:nbn:de:hebis:30-69945

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231-264. doi: 10.1007/s11409-008-9029-x

Enders, C. K., & Bandolos, D. L. (2001). The relative performance to full information maximum likelihood estimation for missing data in structural equation models. . *Structural Equation Modeling, 8*(3), 430–457.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*(0), 1-9. doi: http://dx.doi.org/10.1016/j.learninstruc.2013.07.001

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of Two Anomalous Results in Statistical Mediation Analysis. *Multivariate Behavioral Research, 47*(1), 61-87. doi: 10.1080/00273171.2012.640596

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., . . . Jancek, D. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology, 95*(2), 293-305. doi: 10.1037/0022-0663.95.2.293

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . . Doolittle, F. (2011). Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *Am Psychol, 52*(11), 1182-1186.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. NY: Routledge.

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of Learning Skills Interventions on Student Learning: A Meta-Analysis. *Review of Educational Research, 66*(2), 99-136.

Hayes, A. F., & Scharkow, M. (2013). The Relative Trustworthiness of Inferential Tests of the Indirect Effect in Statistical Mediation Analysis: Does Method Really Matter? *Psychological Science*.

Judd, C. M., & Kenny, D. A. (1981). Process Analysis: Estimating Mediation in Treatment Evaluations. *Evaluation Review, 5*(5), 602-619.

Jonkmann, K., Rose, N., & Trautwein, U. (Eds.) (2013). *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen. Abschlussbericht für die Länder Baden-Württemberg und Sachsen [Tradition and innovation: Academic and psychosocial development in vocational track schools in the states of Baden-Württemberg and Sachsen]*, Unpublished project report, University of Tübingen, Tübingen, Germany .

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A Monte Carlo Comparison Study of the Power of the Analysis of Covariance, Simple Difference, and Residual Change Scores in Testing Two-Wave Data. *Educational and Psychological Measurement, 73*(1), 47-62.

Kistner, S., Rakoczy, K., Otto, B., Dignath-van Ewijk, C., Büttner, G., & Klieme, E. (2010). Promotion of self-regulated learning in classrooms. Investigating frequency, quality, and consequences for student performance. *Metacognition and Learning, 5*(2), 157-171. doi: DOI: 10.1007/s11409-010-9055-3

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". [Quality dimensions and effectiveness of mathematics instruction. Theoretical background and selected findings of the Pythagoras

project]. In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule.* (pp. 127-146). Münster, Westfalen u.a.: Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom.* (pp. 137-160). Münster u.a.: Waxmann.

Kline, F. M., Deshler, D. D., & B.Schumaker, J. (1992). Implementing learning strategy instruction in class settings: A research perspective. In M. Pressley, K. R. Harris & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* . (pp. 361–406). New York: Academic.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel Modeling of Individual and Group Level Mediated Effects. *Multivariate Behavioral Research, 36*(2), 249-277. doi: 10.1207/S15327906MBR3602_06

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527-537. doi: 10.1016/j.learninstruc.2008.11.001

MacKinnon, D. P. (2008). Introduction to Statistical Mediation Analysis. Mahwah, NJ: Erlbaum.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83-104. doi: 10.1037/1082-989X.7.1.83

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data. A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition.*: Los Angeles, CA: Muthén & Muthén.

Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education, 24*(1), 17-31. doi: 10.1007/BF03173472

Perels, F., Gürtler, T., & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction, 15*(2), 123-139. doi: 10.1016/j.learninstruc.2005.04.010

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (Eds.). (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore: Paul H. Brookes.

Pintrich, P. R. (2000). The role ofn goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). San Diego, CA: Academic Press.

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing Between Cross- and Cluster-Level Mediation Processes in the Cluster Randomized Trial. *Sociological Methods & Research, 41*(4), 630-670. doi: 10.1177/0049124112460380

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative Methods for Assessing Mediation in Multilevel Data: The Advantages of Multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 161-182. doi: 10.1080/10705511.2011.557329

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). Befragungsinstrumente. [Questionnaires]. In E. Klieme, C. Pauli & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 1).* Frankfurt a.M.: GFPF/DIPF.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., . . . Schiefele, U. (Eds.). (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente. [Questionnaires]*. Münster: Waxmann.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A Systematic Approach*: SAGE Publications.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*(1), 64-96. doi: http://dx.doi.org/10.1016/j.cedpsych.2005.02.002

Schochet, P. Z. (2008). Guidelines for multiple testing in impact evaluations of educational interventions. Princeton, NJ: Mathematica Policy Research.

Schünemann, N., Spörer, N., & Brunstein, J. C. (2013). Integrating self-regulation in whole-class reciprocal teaching: A moderator–mediator analysis of incremental effects on fifth graders' reading comprehension. *Contemporary Educational Psychology, 38*(4), 289-305. doi: http://dx.doi.org/10.1016/j.cedpsych.2013.06.002

Seidel, T., Rimmele, R., & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction, 15*(6), 539-556. doi: 10.1016/j.learninstruc.2005.08.004

Šidák, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association, 62*(318), 626-633. doi: 10.1080/01621459.1967.10482935

Souvignier, E., & Trenk-Hinterberger, I. (2010). Implementation eines Programms zur Förderung selbstregulierten Lesens. Verbesserung der Nachhaltigkeit durch Auffrischungssitzungen. *Zeitschrift für Pädagogische Psychologie, 24*(3-4), 207-220.

Spörer, N., & Glaser, C. (2010). Förderung selbstregulierten Lernens im schulischen Kontext. *Zeitschrift für Pädagogische Psychologie, 24*(3), 171-175. doi: 10.1024/1010-0652/a000014

Thillmann, H., Künsting, J., Wirth, J., & Leutner, D. (2009). Is it Merely a Question of "What" to Prompt or Also "When" to Prompt? *Zeitschrift für Pädagogische Psychologie, 23*(2), 105-115. doi: 10.1024/1010-0652.23.2.105

Vohs, K. D., & Baumeister, R. F. (Eds.). (2011). *Handbook of self-regulation: Research, theory, and applications* (2. ed.). New York, London: The Guilford Press.

Wagner, W. (2008). Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz. Retrieved from http://kola.opus.hbz-nrw.de/volltexte/2008/234

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*(0), 1-11. doi: http://dx.doi.org/10.1016/j.learninstruc.2013.03.003

Wagner, W., Helmke, A., & Rösner, E. (2009). Deutsch Englisch Schülerleistungen International. Dokumentation der Erhebungsinstrumente für Schülerinnen und Schüler, Eltern und Lehrkräfte. Frankfurt, Main: DIPF.

Wei, R. C., Darling-Hammont, L., Andree, A., Richardson, N., & Orphanos, S. (2009). The status of professional development in the United States. In R. C. Wei, L. Darling-Hammont, A. Andree, N. Richardson & S. Orphanos (Eds.), *Professional learning in the learning profession: A status report on teacher development in the United States and abroad* (pp. 19-27). Standord National Staff Development Council

Weinert, F. E., Schrader, F. W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Research, 13*, 895-914.

Werth, S., Wagner, W., Ogrin, S., Trautwein, U., Friedrich, A., Keller, S., . . . Schmitz, B. (2012). Förderung des selbstregulierten Lernens durch die Lehrkräftefortbildung «Lernen

mit Plan»: Effekte auf fokale Trainingsinhalte und die allgemeine Unterrichtsqualität. [Teaching Teachers how to teach Self-Regulated Learning: Effects of a Training Program on the Promotion of Self-Regulation and Instructional Quality]. *Zeitschrift für Pädagogische Psychologie, 26*(4), 291-305. doi: 10.1024/1010-0652/a000080

Yoon, K. S., Duncan, T., Lee, R. W.-Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement *Issues & Answers Report*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 1–37). Mahwah, NJ: Erlbaum.

5

Gesamtdiskussion

5 Gesamtdiskussion

Ziel der vorliegenden Arbeit war es, erstens eine derzeit noch bestehende Forschungslücke hinsichtlich des Einflusses mehrmaliger Erhebung von Lehrer- und Schülerurteilen zur Erfassung der Unterrichtsqualität zu schließen. Dabei sollte insbesondere untersucht werden, inwieweit sich die im Querschnitt gefundene niedrige bis moderate Übereinstimmung zwischen den beiden Perspektiven durch die Zerlegung in messzeitpunktspezifische und messzeitpunktüberdauernde Varianzanteile verbessern lässt. Zweitens sollte die Arbeit untersuchen, ob im Rahmen einer Lehrerfortbildung zur Förderung des selbstregulierten Lernens der Schülerinnen und Schüler auch eine Veränderung der allgemeinen Unterrichtsqualität—und damit untrainierter Aspekte der Unterrichtspraxis—auftrat. Als dritte Fragestellung wurde in der vorliegenden Arbeit untersucht, ob die in anderen Studien vielfach berichteten Effekte von selbstregulationsbezogenen Lehrerfortbildungsstudien auf die Mathematikleistung auch durch die in Studie 2 gefundenen Veränderungen der allgemeinen Unterrichtsqualität hervorgerufen wurden. Im folgenden Abschnitt sollen zunächst die zentralen Befunde der drei Studien zu den drei dieser Arbeit zugrundeliegenden Fragestellungen zusammengefasst und diskutiert werden. Im Anschluss an die Diskussion der Befunde soll auch auf die Grenzen der vorliegenden Arbeit bzw. der durchgeführten Teilstudien eingegangen werden. Den letzten Abschnitt der vorliegenden Arbeit bildet eine Darstellung möglicher Implikationen für die bereits im Einleitungsteil vorgestellten Forschungsfelder, die den theoretischen Hintergrund der vorliegenden Arbeit bildeten, und für die Praxis.

5.1 Zusammenfassung und Diskussion der zentralen Befunde

5.1.1 Zentrale Befunde zu messzeitpunktspezifischen und messzeitpunktübergreifenden Varianzanateilen und zur Übereinstimmung der messzeitpunktübergreifenden Komponente aus Lehrer- und Schülerperspektive[16]

In Anlehnung an Befunde von Kane und Staiger (2012) zu den Eigenschaften von Beobachterratings zur Erhebung der Unterrichtsqualität wurde im Rahmen von Forschungsfrage 1 der vorliegenden Arbeit untersucht, inwieweit auch Lehrer- und Schülerurteile zur Unterrichtsqualität messzeitpunktespezifische und messzeitpunktüber-

[16] Dieser Abschnitt ist eine wörtliche Übersetzung der Ergebniszusammenfassung aus Studie II, da es aus Sicht der Autorin keine bessere Möglichkeit gibt, die Befunde der Studie stringent zusammenzufassen.

dauernde Varianzanteile aufweisen. Dabei wurde auch untersucht, ob die Lehrer-Schüler-Übereinstimmung für die messzeitpunktübergreifende Komponente höher ist als für die zu einem Messzeitpunkt erzielte Übereinstimmung beider Perspektiven.

Zur Klärung dieser Fragestellung können die Befunde von Studie 1 herangezogen werden. So ließen sich die Lehrer- und Schülerurteile der Unterrichtsqualität in eine messzeitpunktspezifische und eine zeitstabile Komponente zerlegen, wobei dieser Befund den Ergebnissen entspricht, die Kane und Staiger (2012) für Beobachterratings von Unterricht im Rahmen des MET-Projekts ermittelten. Dabei erwies sich insbesondere für die Lehrerratings der Varianzanteil der zeitstabilen Komponente als höher als der entsprechende Varianzanteil in den Beobachterratings, der durch Kane und Staiger (2012) berichtet wurde. Hinsichtlich der Reliabilität der Schülerratings fielen die Ergebnisse in Abhängigkeit der gewählten Ebene unterschiedlich aus. So entfielen fast die ganze Varianz der Schülerratings auf der Klassenebene auf die zeitstabile Komponente. Da jedoch der größte Teil der Varianz nicht auf Klassenebene, sondern innerhalb von Klassen lag, sind die Schülerurteile insgesamt als weniger repräsentativ für die beurteilte (durchschnittliche) Unterrichtspraxis einzuschätzen als die Lehrerurteile. Hinsichtlich des Einflusses der mehrmaligen Erhebung der Unterrichtsqualität lassen sich aus den Ergebnissen von Studie 1 folgende Schlüsse ziehen: Erstens scheinen die Lehrerurteile im Gegensatz zu den Befunden von Koziol und Burns (1986) durch die mehrmalige Erhebung nicht systematisch beeinflusst zu werden. So vergrößerte sich zwar der Varianzanteil der zeitstabilen Komponente in Lehrerurteilen vom ersten zum dritten Messzeitpunkt im Fall von Klassenführung und Autonomieunterstützung; für das Konstrukt Zielsetzung hingegen war der Varianzanteil der zeitstabilen Komponente zum zweiten Messzeitpunkt am größten. Zweitens scheint die mehrmalige Messung auch auf die Schülerratings keinen systematischen Effekt zu haben, da hier der Anteil der zeitstabilen Variante über Zeitpunkte, Konstrukte und Aggregationsebene variierte. So blieb zwar der Varianzanteil der zeitstabilen Komponente auf Klassenebene konstant, jedoch stieg die Varianz innerhalb der Klassen an, weshalb der Anteil der zeitstabilen Komponente im Verhältnis zur angestiegenen Gesamtvarianz abnahm. Folglich kann hier zwar ein Effekt der mehrmaligen Messung auf die Varianz innerhalb von Klassen angenommen werden, nicht jedoch auf den Anteil der zeitstabilen Komponente. Als Erklärung für diesen Anstieg der Varianz innerhalb von Klassen—insbesondere im Fall des Konstrukts Autonomieunterstützung—führt Studie 1 die Möglichkeit an, dass die Schülerinnen und Schüler sich an die Fragen gewöhnt haben und daher ihr Urteil mit der Zeit repräsentativer für

die individuell erlebte Autonomieunterstützung wurde, was dann wiederum zum Anstieg der Varianz innerhalb von Klassen führte.

Ob sich die Übereinstimmung zwischen den Lehrer- und Schülerratings erhöht, wenn man nach Aggregation mehrerer Messzeitpunkte die zeitstabile Komponente beider Perspektiven extrahiert und dann deren Korrelation ermittelt, war ebenfalls Teil der ersten Fragestellung der vorliegenden Arbeit. Die zur Beantwortung dieser Frage ermittelten Ergebnisse aus Studie 1 ergeben auch hier ein heterogenes Bild. So ergab sich durch Extraktion der zeitstabilen Komponente im Fall des Konstrukts Klassenführung eine perfekte Korrelation zwischen den beiden Perspektiven. Auch für das Konstrukt Zielsetzung ergab sich eine leicht höhere Korrelation. Allerdings konnte dies nicht für das Konstrukt Autonomieunterstützung bestätigt werden, da hier die bereits im Querschnitt nicht signifikante Übereinstimmung zwischen Lehrer- und Schülerurteilen auch nach Extraktion der messzeitpunktübergreifenden Komponente nicht signifikant war. Diese Unterschiede zwischen den Konstrukten führt die Studie auf auch zuvor bestehende, konstruktspezifische Unterschiede in der Höhe der Übereinstimmung zwischen den Perspektiven zurück. So war bereits im Querschnitt die Klassenführung das Konstrukt mit der höchsten Lehrer-Schüler-Übereinstimmung und die Autonomieunterstützung das Konstrukt, das auch schon im Querschnitt keine signifikante Korrelation zwischen den beiden Perspektiven aufwies. Die Annahme, dass sich die mangelnde Übereinstimmung womöglich durch die Kontrolle messzeitpunktspezifischer Varianz beheben ließe, musste durch die ausbleibende Verbesserung der Korrelation im Fall von Autonomieunterstützung jedoch verworfen werden. Die Studie schließt daher mit dem Fazit, dass die Kontrolle von messzeitpunktspezifischer Varianz wohl erst zu einer Verbesserung der Übereinstimmung führen kann sobald überhaupt anzunehmen ist, dass die Lehrkräfte und die Schülerinnen und Schüler eine korrespondierende Wahrnehmung des Konstrukts haben. Insbesondere da im Fall von Autonomieunterstützung die Korrelation der Schülerwahrnehmung innerhalb von Klassen über die Messzeitpunkte hinweg abnahm, kann kritisch hinterfragt werden, ob für ein solches Konstrukt überhaupt eine Übereinstimmung zwischen den Perspektiven erwartbar ist. Diese Erklärung ist auch im Einklang mit Rosenshines Taxonomie, gemäß derer Autonomieunterstützung als ein hoch-inferentes Konstrukt zu kategorisieren ist, welches nicht "specific, denotable [and] relatively objective" (1970, S. 281) ist, wohingegen Klassenführung als niedrig-inferentes Konstrukt diese Kriterien erfüllt.

5.1.2 Zentrale Befunde zu den Auswirkungen einer Lehrerfortbildung zur Förderung der Selbstregulation auf die allgemeine Unterrichtsqualität[17]

Fragestellung 2 der vorliegenden Arbeit untersuchte, ob eine Lehrerfortbildung zur Förderung der Selbstregulation im Mathematikunterricht nicht nur Effekte auf die fokalen Trainingsinhalte, sondern auch auf die allgemeine Unterrichtsqualität hatte, deren Verbesserung nicht unmittelbares Ziel des Trainings war.

In Bestätigung der Befunde früherer Studien hatte auch die Teilnahme an der Lehrerfortbildung „Lernen mit Plan“ sowohl aus Lehrer- als auch aus Schülersicht einen positiven Effekt auf die Förderung der Selbstregulation im Unterricht. Die ermittelten Effektstärken umfassten dabei Werte von $0.42 \leq d \leq 0.65$ können damit als bedeutsam eingestuft werden.

Unterschiede zwischen den Lehrer- und Schülerurteilen gab es hingegen im Hinblick auf die untersuchten Aspekte der allgemeinen Unterrichtsqualität. So hatte die Fortbildungsteilnahme aus Sicht der Lehrkräfte im Vergleich zur Kontrollgruppe keinen Effekt auf die Aspekte der allgemeinen Unterrichtsqualität. Im Gegensatz dazu deuteten die Schülerurteile der Experimentalgruppe im Vergleich zur Kontrollgruppe auf einen positiven Effekt auf die allgemeine Unterrichtsqualität hin. Auch dieser, ausschließlich von den Schülerinnen und Schülern berichtete Effekt der Fortbildungsteilnahme, kann mit (zwischen den untersuchten Konstrukten variierenden) Effektstärken von $0.25 \leq d \leq 0.39$ ebenfalls als bedeutsam eingestuft werden. Dabei ist zu erwähnen, dass im Falle der Schülerurteile sowohl der positive Effekt auf die selbstregulationsspezifische als auch der positive Effekt auf die allgemeine Unterrichtsqualität darauf zurückzuführen sind, dass beide Gruppen ein Absinken der Unterrichtsqualität in beiden Bereichen berichten, die Mittelwerte in der Experimentalgruppe jedoch geringer abfallen als in der Kontrollgruppe.

Eine rein inhaltliche Interpretation der Ergebnisse—die u.a. aufgrund der Unterschiede zwischen den Lehrer- und den Schülerurteilen schwierig ist—würde aus Lehrersicht zu der mit früheren Forschungsergebnissen konformen Annahme führen, dass die Teilnahme an der selbstregulationsbezogenen Lehrerfortbildung „Lernen mit Plan“ zu einer Verbesserung der selbstregulationsspezifischen Unterrichtsqualität geführt hat und keine Auswirkungen auf die allgemeine Unterrichtsqualität hatte. Auch aus Schülersicht zeigt sich dieser positive Effekt

[17] Abschnitt 5.1.2 ist entspricht der Ergebniszusammenfassung aus Studie I, da es aus Sicht der Autorin keine bessere Möglichkeit gibt, die Befunde der Studie stringent zusammenzufassen.

der Trainingsteilnahme auf die selbstregulationsspezifische Unterrichtsqualität. Allerdings zeigt sich aus Schülersicht in schwächerer Ausprägung ebenfalls ein positiver Effekt auf die allgemeine Unterrichtsqualität. Wie jedoch das Absinken der Mittelwerte der Schülerratings sowohl der selbstregulationsspezifischen als auch der allgemeinen Unterrichtsqualität über die Zeit zu interpretieren ist, wirft forschungsmethodologische Fragen auf und weist auf einen messbaren Einfluss von Wahrnehmungseffekten auf die Beurteilung von Unterrichtsqualität über die Zeit hin. So deuten die Ergebnisse von Studie 2 darauf hin, dass die Wahrnehmung der Unterrichtveränderung und damit eine Veränderung der Mittelwertstruktur von Messzeitpunkt 1 zu Messzeitpunkt 3 aus Lehrer- und Schülersicht maßgeblich von einander abweicht. Eine mögliche Interpretation dieser gefundenen Unterschiede könnte sein, dass sich die allgemeine Unterrichtsqualität zwar verändert hat, die Lehrkräfte dies aber durch die starke Fokussierung auf die Umsetzung der Fortbildungsinhalte im Gegensatz zur ihren Schülerinnen und Schülern nicht zur Kenntnis genommen haben. Andererseits könnte in Anlehnung an Befunde zur unterschiedlichen Differenziertheit der Lehrer- und Schülerwahrnehmung jedoch auch die These aufgestellt werden, dass die Schülerinnen und Schüler trotz eines ausschließlich spezifischen Fortbildungseffekts auf die selbstregulationsspezifische Unterrichtsqualität auch Veränderungen der allgemeinen Unterrichtsqualität wahrgenommen haben. Unterstützt wird diese weitere Interpretationsmöglichkeit mit dem Befund von Kunter et al. (2005), dass Schülerinnen und Schülern eine globalere Unterrichtswahrnehmung aufweisen als Lehrkräfte. In Anbetracht der unterschiedlichen Stärke der aus Schülersicht berichteten Veränderungen der selbstregulationsspezifischen und allgemeinen Unterrichtsqualität erscheint diese Interpretation jedoch nicht besonders plausibel zu sein. Einen weiteren Ansatz zur Interpretation der Ergebnisse bieten Befunde zu Panel-Conditioning-Effekten im Rahmen der cognitive stimulus-Hypothese (Sturgis, Allum & Brunton-Smith, 2009). So wäre es denkbar, dass die wiederholte Befragung der Schülerinnen und Schüler zu einer genaueren Beobachtung des Unterrichtsgeschehens geführt hat und dieses in Folge strenger beurteilt wurde. Unterstützt wird diese Interpretation durch die über die Messzeitpunkte hinweg absinkenden Mittelwerte der selbstregulationsspezifischen und der allgemeinen Unterrichtsqualität. Das geringere Absinken der Mittelwerte in der Experimentalgruppe könnte diesen Überlegungen zufolge durch eine Überlagerung des positiven Interventionseffekts und eines negativen und für beide Gruppen gleichermaßen zutreffenden „Befragungseffekts“ begründet sein. Die Unterschiede zwischen dem Verlauf der Mittelwertstruktur aus Lehrer- und Schülersicht können dadurch jedoch nicht erklärt werden,

weswegen diese Diskrepanz erneut im Abschnitt zu den Grenzen der vorliegenden Arbeit thematisiert werden soll.

5.1.3 Zentrale Befunde zur Vermittlung des Effekts einer Lehrerfortbildung zur Förderung der Selbstregulation auf die Mathematikleistung von Schülerinnen und Schülern

Als dritte Forschungsfrage untersuchte die vorliegende Arbeit, ob der vielfach beschriebene Effekt von Lehrerfortbildungen zur Förderung der Selbstregulation auf die Mathematikleistung ausschließlich durch die Umsetzung der Trainingsinhalte oder auch über die Veränderung der allgemeinen Unterrichtsqualität vermittelt wird. Die Ergebnisse von Studie 3 deuten dabei darauf hin, dass auch die Lehrerfortbildung „Lernen mit Plan“ einen positiven Effekt auf die Mathematikleistung der Schülerinnen und Schüler hatte. Die durchgeführten Mediationsanalysen zeigen dabei, dass die Vermittlung dieses Effekts durch die Umsetzung der fokalen Trainingsinhalte erfolgte. Auch hinsichtlich der Vermittlung des Fortbildungseffekts durch die Unterrichtsqualität fand sich ein leichter Effekt, der jedoch nach Adjustierung für multiples Testen nicht mehr signifikant war.

Studie 3 greift damit die Ergebnisse von Studie 2 auf, die zeigen, dass sich die Lehrerfortbildung „Lernen mit Plan“ aus Sicht der Schülerinnen und Schüler auch auf die allgemeine Unterrichtsqualität ausgewirkt hat, ohne dass diese durch das Training gezielt adressiert wurde. Auch wenn dieser Effekt der Lehrerfortbildung—wie in Studie 2 gezeigt—nur aus Sicht der Schülerinnen und Schüler eingetreten ist, so gab dieser Befund doch Anlass zu den weiterführenden Mediationsanalysen in Studie 3, in denen untersucht wurde, ob die aus Schülersicht eingetretene Veränderung der allgemeinen Unterrichtsqualität auch einen Einfluss auf die Veränderung der Mathematikleistung der Schülerinnen und Schüler hatte. Anlass zu dieser weiteren Untersuchung boten dabei zweierlei Forschungsbefunde: Zum einen wird die von Schülerinnen und Schülern eingeschätzte Unterrichtsqualität als prädiktiver für die Schülerleistung eingestuft als die durch Lehrkräfte und Beobachter eingeschätzte Unterrichtsqualität. Zum anderen stellen Spörer und Glaser (2010) fest, dass die erfolgreiche Implementation von Fortbildungsinhalten—sofern diese überhaupt überprüft wird—häufig nicht zu den erwarteten Effekten auf die Leistung der Schülerinnen und Schüler führt.

Im Fall der Lehrerfortbildung „Lernen mit Plan“ zeigte sich, dass sich entsprechend der Forschungsliteratur ein Effekt der Fortbildungsteilnahme auf die Mathematikleistung der Schülerinnen und Schüler eingestellt hat. Darüber hinaus zeigte sich, dass dieser

Fortbildungseffekt erwartungskonform über die Veränderung der selbstregulations-spezifischen Unterrichtsqualität vermittelt wurde. Jedoch weisen die Befunde darauf hin, dass der Effekt der Fortbildung auf die Mathematikleistung erstens nicht vollständig über die Veränderung der selbstregulationsspezifischen Unterrichtsqualität vermittelt wurde. Zweitens war zunächst auch eine Vermittlung des Fortbildungseffekts auf die Mathematikleistung durch die Veränderung der allgemeinen Unterrichtsqualität feststellbar, die jedoch nach der Korrektur für multiples Testen nicht mehr signifikant war. Setzt man diesen Befund in Zusammenhang mit den Ausführungen von Spörer und Glaser (2010), so könnte dies ein Hinweis darauf sein, dass sich auch bei anderen selbstregulationsfördernden Lehrerfortbildungen möglicherweise Veränderungen der allgemeinen Unterrichtsqualität eingestellt haben, die in diesen Studien jedoch nicht erfasst wurden und deren Einfluss auf die Mathematikleistung daher nicht untersucht werden konnte. Sollte sich in diesen Studien jedoch wie im Theorieteil der vorliegenden Arbeit ausgeführt möglicherweise ein negativer Effekt auf die allgemeine Unterrichtsqualität eingestellt haben, so könnte dies dazu führen, dass die antizipierten positiven Effekte auf die Schülerleistung geringer ausfallen als erwartet. Studie 3 gibt damit nicht nur einen Einblick in die Wirkmechanismen von selbstregulationsfördernden Fortbildungen, sondern auch einen Hinweis darauf, welche möglichen Auswirkungen die Integration neuer Unterrichtspraxis auf das Gesamtgefüge der Unterrichtsgestaltung und damit die Leistungsentwicklung der Schülerinnen und Schüler hat.

5.2 Grenzen der vorliegenden Arbeit

In diesem Abschnitt soll auf die Grenzen eingegangen werden, die sich aus dem Design der zugrundeliegenden Lehrerfortbildung "Lernen mit Plan" für die Klärung der in dieser Dissertation behandelten Forschungsfragen ergeben haben. Weitere Einschränkungen, die sich bei der Klärung der spezifischen Fragen der Teilstudien ergeben haben und dort intensiv erörtert wurden, sollen an dieser Stelle nicht diskutiert werden.

So unterliegt auch die Lehrerfortbildungsstudie "Lernen mit Plan" der Problematik, dass im Rahmen von Interventionsstudien aufgrund des großen Aufwands häufig nur relativ kleine Stichprobengrößen erzielt werden können. Beispielsweise basieren die in der vorliegenden Arbeit erwähnten selbstregulationsfördernden Interventionsstudien nur auf Stichproben zwischen 2 (Perels, Dignath & Schmitz, 2009) und 27 (Souvignier & Trenk-Hinterberger, 2010) Klassen Im Gegensatz dazu ist die Studie „Lernen mit Plan" mit einer realisierten Stichprobengröße von 78 Klassen nicht nur im Vergleich zu Interventionsstudien im Bereich der Selbstregulationsförderung, sondern auch im Vergleich zu anderen

Lehrerfortbildungsstudien als relativ große Studie anzusehen (Yoon, Duncan, Lee, Scarloss & Shapley, 2007). So wiesen Yoon et al. (2007) in einer Metaanalyse zum Einfluss von Lehrerfortbildungen auf Schülerleistungen darauf hin, dass unter 1343 Studien lediglich neun über ein quasiexperimentelles Design oder ein randomisiertes Kontrollgruppendesign verfügten. Diese wiederum umfassten Stichprobengrößen von lediglich 5 bis 44 Lehrkräften, was die Generalisierbarkeit der Befunde dieser Studien einschränkt. Dennoch reicht selbst die vergleichsweise große Stichprobe der Studie „Lernen mit Plan“ nicht aus, um komplexe Mehrebenenmodelle zu spezifizieren und zu analysieren. Insbesondere bei Analysen im Bereich der Unterrichtsqualität, die nach dem aktuellen Stand der Forschung durch aggregierte Schülerurteile auf der Klassenebene ermittelt wird (Lüdtke, Robitzsch, Trautwein & Kunter, 2009), war es in den durchgeführten Analysen beispielsweise nicht möglich, die Skalen latent zu bilden und das zugrundeliegende Messmodell zu überprüfen. Messfehler, die sich möglicherweise durch die Qualität der einzelnen Items ergeben haben, konnten damit in keiner der durchgeführten Teilstudien kontrolliert werden.

Als weitere Grenze der Arbeit sind Einschränkungen in der Auswahl der verwendeten Skalen zu nennen. Auch hier unterliegen alle Teilstudien dem Problem, dass im Rahmen der zur Klärung zahlreicher Forschungsfragen konzipierten Lehrerfortbildungsstudie „Lernen mit Plan“ nur eine eingeschränkte Anzahl an Skalen zur Erfassung der Unterrichtsqualität in den Lehrer- und Schülerfragebögen eingesetzt werden konnten. So musste in Anbetracht der drei Erhebungszeitpunkte streng darauf geachtet werden, dass die Fragebögen nicht zu umfangreich wurden. Dieses Problem wurde noch weiter verschärft durch das Alter und die Schulformzugehörigkeit der Schülerinnen und Schüler. So gibt es für die Zielgruppe von Hauptschülern der fünften Jahrgangsstufe nur wenige empirisch bewährte Skalen zur Erfassung von Unterrichtsqualität. Dem Anspruch folgend, ausschließlich bewährte Skalen einzusetzen und gleichzeitig die Parallelität der erfragten Konstrukte bei Lehrkräften und Schülerinnen und Schülern zu gewährleisten, konnten feinere Unterschiede zwischen den eingesetzten Instrumenten wie beispielsweise der in der Unterrichtsforschung häufig thematisierte Inferenzgrad in den Analysen nicht berücksichtig werden.

Auch führte die Zielgruppe von Schülerinnen und Schülern der fünften Jahrgangsstufe der Schulform Hauptschule zu einer limitierten Qualität der Schülerdaten. So genügen die eingesetzten Lehrer- und Schülerskalen zwar den üblichen Kriterien wie bspw. Cronbach’s alpha. Trotzdem musste festgestellt werden, dass die Interklassenkorrelation zwar akzeptabel aber dennoch niedriger war, als dies für die Spezifikation anspruchsvollerer Analysen notwendig gewesen wäre.

Eine weitere Grenze der Studie „Lernen mit Plan" stellt der durch limitierte Ressourcen begründete Verzicht auf Beobachterratings in Ergänzung zu den Lehrer- und Schülerurteilen zur Erfassung der Unterrichtsqualität dar. In Anbetracht der Feststellung von Clausen (2002), dass Lehrer-, Schüler- und Beobachterratings spezifische Wahrnehmungsperspektiven darstellen und Beobachterratings weniger anfällig für eine Verzerrung durch selbstdienliche Wahrnehmungstendenzen oder durch Halo-Effekte sind, wären Beobachterratings insbesondere zur Interpretation der Veränderung der Mittelwertstruktur der Unterrichtsqualität im Verlauf der Intervention hilfreich gewesen. So wurde in Teilstudie 2 auf den ungewöhnlichen Befund der Studie verwiesen, dass sich die positiven Effekte der Trainingsteilnahme auf die untersuchten Outcomes darüber ergaben, dass sich bei den Schülerinstrumenten sowohl für die Experimentalgruppe als auch für die Kontrollgruppe ein Absinken der avisierten Outcomes, selbstregulationsspezifische und allgemeine Unterrichtsqualität, fand, das Absinken in der Experimentalgruppe jedoch geringer ausfiel als in der Kontrollgruppe. Zwar konnte aufgrund des für Lehrerfortbildungen außergewöhnlichen randomisierten Kontrollgruppendesigns trotz dieses Absinkens der positive Effekt der Trainingsteilnahme festgestellt werden. Dennoch ist es aufgrund der fehlenden Beobachterratings nicht möglich, ein tatsächliches Absinken der Unterrichtsqualität auszuschließen und das Absinken stattdessen auf Befragungseffekte auf Seiten der Schülerinnen und Schüler zurückzuführen. Auch wenn Studie 2—insbesondere in Anbetracht der von den Schülerurteilen abweichenden Lehrerurteile—die Möglichkeit von Befragungseffekten durch die wiederholte Befragung intensiv erörtert, kann diese Frage in Ermangelung von Beobachterratings und durch einen Mangel an Hinweisen auf dieses Phänomen in anderen Forschungsarbeiten nicht abschließend geklärt werden.

5.3 Implikationen für die Forschung

In diesem Abschnitt soll vor allem auf Implikationen für die Forschung in den Bereichen der Erfassung von Unterrichtsqualität und der Evaluation von Lehrerfortbildungen eingegangen werden, die in den einzelnen Teilstudien noch nicht intensiv thematisiert wurden und sich vor allem aus der Gesamtschau der vorliegenden Dissertation ergeben.

5.3.1 Implikationen für die Unterrichtsforschung

Die Studie „Lernen mit Plan" kombinierte eine intensive Lehrerfortbildung mit einem methodisch anspruchsvollen Wartekontrollgruppendesign und einer im Verhältnis zu anderen Interventionsstudien umfangreichen Stichprobe von 78 Klassen. Basierend auf diesem Datensatz war es möglich, Forschungsfragen im Bereich der Unterrichtsqualität nachzugehen, die in dieser Form noch nicht untersucht werden konnten. So gibt es zu der Frage, wie sich in Lehrer- und Schülerwahrnehmungen variable und konstante Aspekte der Unterrichtsqualität wiederspiegeln, nach Kenntnisstand der Autorin, bisher keine empirischen Befunde. Auch die Frage, inwieweit die Übereinstimmung von Lehrer- und Schülerwahrnehmung durch diese situationsspezifischen Varianzanteile beeinflusst wird, wurde bisher empirisch nicht untersucht. In Anbetracht der Tatsache, dass zur Erfassung von Unterrichtsqualität aufgrund der geringeren Kosten Lehrer- und Schülerratings deutlich häufiger eingesetzt werden als Beobachterratings (Desimone, Smith & Frisvold, 2009), ist es von großer Bedeutung die Eigenschaften dieser Instrumente besser zu untersuchen. So deutete Clausen (2002) mit einem Hinweis auf Weinstein (1985) zwar darauf hin, dass Schülerurteile möglicherweise messzeitpunktspezifischen Einflüssen unterworfen sind, eine empirische Überprüfung stand jedoch bisher noch aus. Im Hinblick auf die Befunde von Teilstudie 1 lässt sich feststellen, dass nicht nur—wie bereits in anderen Studien gezeigt (z.B. Clausen, 2002; Desimone et al., 2009; Kunter & Baumert, 2006)—die gewählte Perspektive einen Einfluss auf die Bewertung des Unterrichts hat, sondern auch der jeweilige Messzeitpunkt eine nicht zu unterschätzende Rolle spielt. Obwohl bereits frühere Forschungsbefunde darauf hindeuten, dass die Unterrichtsgestaltung im Verlauf der Zeit variiert (z.B. Pianta, Belsky, Vandergrift, Houts & Morrison, 2008; Seidel, Rimmele & Prenzel, 2005) und sich diese Variabilität auch in Beobachterratings wiederspiegelt (Kane & Staiger, 2012), wurden diese Überlegungen bisher noch nicht auf Lehrer- und Schülerfragebögen als Instrumente zur Erfassung von Unterrichtsqualität übertragen. Als Implikation für die weitere Unterrichtsforschung folgt daraus, dass besonders im Zuge des Aufbaus eines breitausgebauten Monitorings des Bildungssystems noch einmal stärker beleuchtet werden sollte, wie die eingesetzten Instrumente funktionieren und welchen Einfluss Befragungseffekte auf die häufig anstelle von Beobachterratings eingesetzten Lehrer- und Schülerinstrumente haben. So hat Teilstudie 2 deutlich herausgearbeitet, dass—unabhängig vom positiven Effekt der Trainingsteilnahme auf die Unterrichtsqualität—die tatsächliche Veränderung der Unterrichtsqualität einen negativen Verlauf hatte, der jedoch in der Kontrollgruppe negativer ausfiel als in der

Experimentalgruppe. Ob sich dieser Befund durch eine tatsächliche Verschlechterung der Unterrichtsqualität oder aber durch eine Überlagerung des Interventionseffekts mit einem möglichen Befragungseffekt erklären lässt, muss von zukünftigen Forschungsarbeiten geklärt werden. Ungeachtet dieser noch offenen Fragen leistet die vorliegende Dissertation damit einen wertvollen Beitrag zur Erforschung der Unterrichtsqualität bzw. deren Erfassung durch Lehrer- und Schülerfragebögen.

5.3.2 Implikationen für die Erforschung von Lehrerfortbildungen zur Förderung der Selbstregulation

Die Studie "Lernen mit Plan" wurde konzipiert, um die Selbstregulation von Fünftklässlern im Mathematikunterricht zu verbessern. Wie in den Teilstudien 2 und 3 gezeigt, war die Lehrerfortbildung ausreichend intensiv konzipiert, um die Unterrichtpraxis der Lehrkräfte im Hinblick auf die fokalen Trainingsinhalte sowohl aus Sicht der Lehrkräfte als auch aus Sicht der Schülerinnen und Schüler zu verändern. Außerdem wurde in dieser Studie gezeigt, dass sich durch die Trainingsteilnahme die Mathematikleistung der Schülerinnen und Schüler veränderte. Dieser Befund ist zwar ebenfalls erwartungskonform, ein Novum ist dabei jedoch, dass die Studie die tatsächliche Vermittlung der Fortbildungseffekte auf die Mathematikleistung durch die Umsetzung der Trainingsinhalte in den Unterricht belegen konnte. Da dieser Nachweis in Teilstudie 3 jedoch nur anhand von den für die Schülerleistung als prädiktiver geltenden Schülerurteilen geführt wurde, steht ein Beleg für diesen Mechanismus anhand von Lehrerurteilen noch aus. Dennoch kann festgestellt werden, dass durch die empirische Überprüfung dieses Mediationsprozesses die durch Spörer und Glaser (2010) bemängelte Forschungslücke zur Wirkung von Lehrerfortbildungen zur Förderung der Selbstregulation von Schülerinnen und Schülern auf deren Leistungsentwicklung geschlossen wurde.

Darüber hinaus konnte basierend auf den Daten der zugrundeliegenden Studie gezeigt werden, dass sich eine so intensiv wie "Lernen mit Plan" konzipierte Fortbildung nicht nur auf die Umsetzung der Trainingsinhalte, sondern auch auf andere, nicht trainierte Aspekte der Unterrichtsqualität auswirkt.

Besondere Bedeutung hat dieser Befund durch die große Bedeutung der Unterrichtsqualität für die Leistungsentwicklung von Schülerinnen und Schülern (Klieme, Pauli & Reusser, 2009), die von zahlreichen Interventionsstudien zur Förderung der Selbstregulation als zentrales Outcome zur Ermittlung der Interventionseffekte herangezogen wird (Dignath & Büttner, 2008). Die in Teilstudie 3 durchgeführte Überprüfung, ob diese

bisher nicht untersuchten Effekte auf die allgemeine Unterrichtsqualität auch die Fortbildungseffekte auf die Mathematikleistung mediieren, ergab jedoch, dass dieser vermutete Mediationseffekt nicht signifkant war. Dennoch lässt die ermittelte unvollständige Mediation darauf schließen, dass die Vermittlung der Effekte auf die Mathematikleistung nicht ausschließlich auf die Umsetzung der Trainingsinhalte zurückzuführen ist. Als Implikation für die Erforschung der Mechanismen von Lehrerfortbildungen zur Förderung der Selbstregulation der Schülerinnen und Schüler im Unterricht ergibt sich aus der Beantwortung der Forschungfragen von Teilstudie 2 und 3 das Plädoyer, die Umsetzungprozesse von Fortbildungsinhalten in die Unterrichtspraxis in zukünftigen Studien verstärkt in den Blick zu nehmen. Eine weitere Implikation, die sich aus Teilstudie 3 ergibt, ist die Klärung der Frage, ob sich die von Dignath und Büttner (2008) ermittelten schwächeren Effekte für Lehrerfortbildungen im Vergleich zu Trainings durch geschulte Assistenten durch mögliche Nebenwirkungen auf die allgemeine Unterrichtsqualität ergeben haben. Nachdem dieser vermutete Effekt nicht bestätigt werden konnte, muss davon ausgegangen werden, dass sich die schwächeren Effekte auf die Schülerleistung eher durch die im Vergleich zu geschulten Assitenten schlechtere Umsetzung der Fortbildungsinhalte durch Lehrkräfte erklären lassen.

Anhand der Befunde der Lehrerfortbildungsstudie “Lernen mit Plan” können jedoch nicht nur die in der vorliegenden Arbeit behandelten Forschungsfragen näher untersucht und geklärt werden. Als groß angelegte Interventionsstudie, die zur Verbesserung der Selbstregulation von Schülerinnenen und Schülern 78 Lehrkräfte fortbildete, ergeben sich außerdem Implikationen für die Lehrerfortbildungsforschung im Allgemeinen. So lässt sich aufgrund der vorliegenden Befunde fragen, inwieweit der Aspekt von möglichen „Nebenwirkungen“ bzw. „Spillover“-Effekten (Desimone, Porter, Garet, Yoon & Birman, 2002) nicht stärker in den Blick genommen werden müsste. Zwar behandelt die Forschungsliteratur zu Interventionsstudien im pädagogisch-psychologischen Bereich durchaus die Thematik von möglichen „Nebenwirkungen“ (Hager & Hasselhorn, 2008), eine systematische Aufarbeitung dieser Thematik findet jedoch nach Kenntnis der Autorin im Rahmen von Lehrerfortbildungsstudien bisher nicht statt. Zur Erlangung eines besseren Verständnisses der Mechanismen von Lehrerfortbildungen (zur Förderung der Selbstregulation) ist es daher wichtig, nicht nur die von Kirkpatrick (1996) beschriebenen vier Ebenen der Trainingsevaluation zu untersuchen, sondern auch zu überlegen, welche potenziellen Nebenwirkungen auf den verschiedenen Ebenen auftreten könnten und wie diese am besten erfasst werden können. Im Falle der Studie „Lernen mit Plan“ wurden die potenziellen Nebenwirkungen auf Ebene 3 des Modells von Kirkpatrick, d.h. der Umsetzung der

Fortbildungsinhalte im Unterricht, erfasst, da es orientiert an Befunden von Borko und Putnam (1995), Boshuizen (2004), Bromme (1992) und Berliner (2001) am wahrscheinlichsten erschien, dass die Nebenwirkungen auf dieser Ebene zu erwarten sind. Da sich in der Lehrerfortbildung "Lernen mit Plan" die Annahme möglicher Nebenwirkungen auf die Unterrichtspraxis bestätigen ließ, wäre es nun wichtig, im Rahmen zukünftiger Lehrerfortbildungsstudien zu überprüfen, inwieweit auch andere Lehrerfortbildungen von vergleichbarer Intensität zu einer Veränderung nicht trainierter Aspekte der Unterrichtsqualität führen und diese Veränderungen möglicherweise auch einen Einfluss auf die Leistungsentwicklung der Schülerinnen und Schüler haben.

Eine weitere Implikation für die Lehrerfortbildungsforschung ergibt sich aus dem Befund von "Lernen mit Plan", dass sich die Lehrerkräfte in ihrer Einschätzung der Veränderung der selbstregulationsspezifischen und der allgemeinen Unterrichtsqualität von den Einschätzungen der Schülerinnen und Schüler unterscheiden. Auch hier geht zwar insbesondere Teilstudie 2 auf das mögliche Vorliegen von Befragungseffekten ein. Da jedoch aufgrund der geringen Stichprobengrößen von Lehrerfortbildungsstudien häufig Lehrer- und Schülerurteile zu den gleichen Konstrukten nicht gegenübergestellt werden können, gibt es nach Kenntnisstand der Autorin keine Lehrerfortbildungsstudie, welche die unterschiedliche Entwicklung der Mittelwertsstruktur von Lehrer- und Schülerurteilen über verschiedene Messzeitpunkte thematisiert und interpretiert. Davon ausgehend, dass sich die Evaluation von Lehrerfortbildungen, die das Ziel einer Unterrichtsveränderung haben, häufig auf die Auswertung von Schülerurteilen beschränkt, wäre in Anbetracht der vorgestellten Befunde dringend zu empfehlen, in zukünftigen Studien durch größere Stichproben auch den Einbezug von Lehrerratings zu ermöglichen. Darüber hinaus wäre es in Anbetracht des Mangels an Befunden zur Übereinstimmung von Lehrer- und Schülerwahrnehmungen der Unterrichtsqualität im Rahmen von Interventionsstudien außerdem wichtig, neben Lehrer- und Schülerratings auch Beobachterratings zu erheben, um weiterführende Erkenntnisse zu spezifischen Einflüssen von Lehrerfortbildungen auf die Unterrichtswahrnehmung von Lehrkräften und Schülerinnen und Schülern zu gewinnen (Clausen, 2002; Wagner, 2008).

5.4 Implikationen für die Praxis

Ebenso wie für die Unterrichtsforschung und die Forschung zu selbstregulations-fördernden Lehrerfortbildungen ergeben sich aus der vorliegenden Arbeit auch für den Unterricht und die Konzeption von Lehrerfortbildungen einige Implikationen.

5.4.1 Implikationen für den Unterricht

Im Hinblick auf die tägliche Unterrichtsgestaltung von Lehrkräften lässt sich aus den Befunden der drei Teilstudien schließen, dass Schülerinnen und Schüler Unterricht nicht nur bewusst erleben, sondern auch Veränderungen wahrnehmen. Ungeachtet der in der vorliegenden Arbeit nicht abschließend geklärten Frage, ob die Unterschiede zwischen der Lehrer- und der Schülerwahrnehmung auf Befragungseffekte zurückzuführen sind, gibt dieser Befund Anlass dazu, Lehrkräften zu empfehlen, die Rückmeldung von Schülerinnen und Schülern zur Unterrichtsgestaltung verstärkt einzuholen. Nicht zuletzt da sich Wahrnehmung von Schülerinnen und Schülern als wichtiger Prädiktor von Leistungsentwicklung erwiesen hat (Klieme et al., 2009), könnten Lehrkräfte dieses Feedback zur Unterrichtsentwicklung nutzen (Marsh & Roche, 1997). Die dabei in allen Teilstudien festgestellten Abweichungen zwischen Lehrer- und Schülerurteilen könnten dabei durch den regelmäßigen gemeinsamen Austausch über das Unterrichtsgeschehen systematisch verringert werden. So wurde insbesondere in Teilstudie 1 darauf hingewiesen, dass sich die Übereinstimmung von Lehrer- und Schülerurteilen unter Kontrolle situationsspezifischer Einflüsse verringert und im Falle von Klassenführung sogar eine vollkommene Übereinstimmung vorliegt. Sollten daher Lehrkräfte und Schülerinnen bzw. Schüler im Rahmen eines regelmäßigen Austausch lernen, in ihrer Rückmeldung einen stärkeren Fokus auf die durchschnittliche Unterrichtsqualität zu legen, könnte selbst mit sehr jungen Klassenstufen bereits ein „Austausch auf Augenhöhe“ stattfinden. Ein weiterer Vorteil eines solchen Austauschs wäre, auch die Differenzen in der Wahrnehmung der Schülerinnen Schüler zu thematisieren und möglicherweise zu verringern. Im Falle von Unterrichtsaspekten, deren Wahrnehmung innerhalb von Klassen stark variiert und die tatsächlich nicht nur unterschiedlich beurteilt, sondern auch unterschiedlich erlebt werden (z.B. Verständlichkeit), könnte ein solcher Dialog der Lehrkraft helfen, die unterschiedlichen Bedürfnisse besser kennen zu lernen und zu berücksichtigen.

5.4.2 Implikationen für die Forschung zu Lehrerfortbildungen zur Förderung der Selbstregulation im Unterricht

In Anbetracht der sehr ernüchternden Befunde von Coleman et al. (1966) und Jencks (1973), dass das Lernen von Schülerinnen und Schüler zu einem deutlich geringeren Prozentsatz als erwartet durch Unterricht und Lehrerhandeln zu beeinflussen ist, ist es doch erfreulich zu sehen, dass wenigstens das Lehrerhandeln durch die Fortbildung von Lehrkräften aktiv gestaltet werden kann.

So zeigt die vorliegende Arbeit, dass Lehrerfortbildungen ein geeignetes Mittel sind, um die Selbstregulation von Schülerinnen und Schülern zu fördern. In Anbetracht der Feststellung von Dignath und Büttner (2008), dass der Einsatz von geschulten Assistenten zu größeren Effekten auf die Schülerinnen und Schülerinnen führt als die Schulung von Lehrkräften, ist dieser Befund ermutigend für die Durchführung von Lehrefortbildungen. So deutet der Befund von Teilstudie 2, der zeigt, dass die sich die Effekte der Fortbildungsteilnahme stärker zum dritten Messzeitpunkt zeigten, darauf hin, dass sich die Effekte vor allem in der langfristigen Umsetzung zeigen. Dies bedeutet zum Einen, dass die schwächeren Effekte von Lehrerfortbildungen im Vergleich zu Trainings durch geschulte Assistenten möglicherweise davon herrühren, dass die Effekte der Interventionen in der Metaanalyse von Dignath und Büttner (2008) einen zu frühen Messzeitpunkt für die Evaluation der Effekte gewählt hatten. Zum Anderen bedeutet dies für die Durchführung von Lehrerfortbildungen zur Förderung von Selbstregulation im Unterricht, dass diese Fortbildungen langfristig angelegt sein sollten, um später die gewünschte Nachhaltigkeit zu erzielen. Hierbei besteht die große Chance, dass die geschulte Lehrkraft nicht nur die Vermittlung der Strategien an die Schülerinnen und Schüler über einen längeren Zeitpunkt umsetzen kann, sondern auch, dass die Lehrkraft zusätzlich die Rolle als Modell für Selbstregulation im Mathematikunterricht aktiv wahrnehmen kann.

Eine weitere Implikation für die Durchführung von Lehrerfortbildungen zur Förderung von Selbstregulation im Unterricht ergibt sich aus dem Befund, dass sich im Rahmen der Studie nicht nur eine Veränderung der fokalen Trainingsinhalte eingestellt hat, sondern auch eine Veränderung der allgemeinen Unterrichtsqualität. Zwar wurde die Veränderung der Mathematikleistung nicht durch diese unintendierte Veränderung vermittelt—dennoch sollte in zukünftigen Lehrerfortbildungen zur Förderung der Selbstregulation intensiv darauf geachtet werden, wie das jeweilige Training möglicherweise die allgemeine Unterrichtsqualität beeinflussen könnte und wie man mögliche negative Effekte dieser Veränderung auf die Mathematikleistung unterbinden bzw. unterstützende Effekte fördern kann.

5.5 Literatur

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463-482. doi: http://dx.doi.org/10.1016/S0883-0355(02)00004-6

Borko, H. & Putnam, R. T. (1995). Expanding a teachers' knowledge base: A cognitive psychologica perspective on professional development. In T. R. Guskey & M. Huberman (Hrsg.), *Professional development in education: New paradigms and practices* (35–66). New York: Teachers College Press.

Boshuizen, H. P. A. (2004). Does practice make perfect? A slow and discontinuous process. In H. P. A. Boshuizen, R. Bromme & H. Gruber (Hrsg.), *Professional learning: Gaps and transitions on the way from novice to expert* (3-8). Dordrecht: Kluwer Academic Press.

Bromme, R. (1992). *Der Lehrer als Experte : zur Psychologie des professionellen Wissens*. Bern {[u.a.]: Huber.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt - und Kriteriumsvalidität. [Quality of instruction: A matter of perspective?]*. Münster, Westfalen u.a.: Waxmann.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. et al. (1966). *Equality of educational opportunity*. Washington, DC US Government Printing Office.

Desimone, L., Porter, A. C., Garet, M. S., Yoon, K. S. & Birman, B. F. (2002). Effects of Professional Development on Teachers' Instruction: Results from a Three-year Longitudinal Study. *Educational Evaluation and Policy Analysis, 24*(2), 81-112. doi: 10.3102/01623737024002081

Desimone, L., Smith, T. & Frisvold, D. (2009). Survey measures of classroom Instruction: Comparing student and teacher reports. *Educational Policy*. doi: 10.1177/0895904808330173

Dignath, C. & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning, 3*(3), 231-264. doi: 10.1007/s11409-008-9029-x

Hager, W. & Hasselhorn, M. (2008). Pädagogisch-psychologische Interventionsmaßnahmen. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (339-347). Göttingen: Hogrefe.

Jencks, C. (1973). *Chancengleichheit*. Reinbek: Rowohlt.

Kane, T. J. & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. In B. M. G. Foundation (Hrsg.), *MET Project Research Paper*.

Kirkpatrick, D. (1996). Great Ideas Revisited. *Training & Development, 50*(1), 6.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom.* (137-160). Münster u.a.: Waxmann.

Koziol, S. M. & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Educational Research, 79*(4).

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231-251. doi: 10.1007/s10984-006-9015-7

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W. et al. (2005). Der Mathematikunterricht der PISA- Schülerinnen und Schüler. *Zeitschrift für Erziehungswissenschaft, 8*(4), 502-520. doi: 10.1007/s11618-005-0156-8

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131. doi: 10.1016/j.cedpsych.2008.12.001

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Perels, F., Dignath, C. & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education, 24*(1), 17-31. doi: 10.1007/BF03173472

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R. & Morrison, F. J. (2008). Classroom Effects on Children's Achievement Trajectories in Elementary School. *American Educational Research Journal, 45*(2), 365-397. doi: 10.3102/0002831207308230

Rosenshine, B. (1970). Evaluation of Classroom Instruction. *Review of Educational Research, 40*(2), 279-300. doi: 10.3102/00346543040002279

Seidel, T., Rimmele, R. & Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and Instruction, 15*(6), 539-556. doi: 10.1016/j.learninstruc.2005.08.004

Souvignier, E. & Trenk-Hinterberger, I. (2010). Implementation eines Programms zur Förderung selbstregulierten Lesens. Verbesserung der Nachhaltigkeit durch Auffrischungssitzungen. *Zeitschrift für Pädagogische Psychologie, 24*(3-4), 207-220.

Spörer, N. & Glaser, C. (2010). Förderung selbstregulierten Lernens im schulischen Kontext. *Zeitschrift für Pädagogische Psychologie, 24*(3), 171-175. doi: 10.1024/1010-0652/a000014

Sturgis, P., Allum, N. & Brunton-Smith, I. (2009). Attitudes Over Time: The Psychology of Panel Conditioning *Methodology of Longitudinal Surveys* (113-126): John Wiley & Sons, Ltd.

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz.* Retrieved from http://kola.opus.hbz-nrw.de/volltexte/2008/234

Weinstein, R. S. (1985). Student mediation of classroom expectancy effects. In J. B. Dusek (Hrsg.), *Teacher expectancies.* Hillsdale, NJ: Erlbaum.

Yoon, K. S., Duncan, T., Lee, R. W.-Y., Scarloss, B. & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement *Issues & Answers Report*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.