Formative Beurteilung im jahrgangsübergreifenden Unterricht

Robbert Smit, Eva Engeli

Zusammenfassung Aus Gründen der Schülerzahlen wird in kleineren Schulen öfters jahrgangsübergreifender Unterricht durchgeführt. In jüngster Zeit greifen aber aus pädagogischen Motiven auch grössere Schulen zu diesem Unterrichtskonzept, welches prädestiniert für den Einsatz einer alternativen, erweiterten Beurteilung scheint: In jahrgangsgemischten Klassen verwendet die Lehrperson vermehrt auf formativen Beurteilungen basierende individuelle und differenzierende Lernziele. Die Frage ist, ob sich Lehrpersonen mit jahrgangsgemischten Klassen bezüglich der Nutzung formativer Beurteilung unterscheiden. Im Rahmen des Projektes „Schule im alpinen Raum 2“ erhoben wir Daten von 280 freiwillig teilnehmenden Lehrpersonen aus zwei Regionen der Ostschweiz und einer in Westösterreich. Unser MixedMethods Forschungsdesign war querschnittlich angelegt. Mit Hilfe eines Strukturgleichungsmodells kann gezeigt werden, dass Lehrpersonen mit jahrgangsgemischten Klassen eine Kombination von Elementen der formativen Beurteilung einsetzen, um den Unterricht zu individualisieren und zu differenzieren. Je höher die Anzahl Jahrgangsstufen sind und je höher die Selbsteinschätzung der Diagnosekompetenz ist, desto häufiger wird eine formative Beurteilung verwendet. Als Ergebnis von zwei Latent Class Analysen zeigen sich zwei Gruppen unterschiedlicher Beurteilungskompetenz, von denen eher wenige Lehrpersonen zur Gruppe „erweiterte Beurteilung“ gehören. In der Diskussion werden die Ergebnisse für Ausbildung und Forschung diskutiert

Schlüsselwörter: Formative Beurteilung; Jahrgangsübergreifender Unterricht; Quantitative und qualitative Methoden; Ländliche Schulen

The role of formative assessment in mixed-age teaching

Abstract Mixed-age teaching often occurs in small, rural schools due to small numbers of students. Recently, however, larger schools are also choosing this teaching configuration for pedagogical reasons. This teaching arrangement necessitates an expanded view of assessment. In multi-grade classes, the teacher focuses on individually-based formative assessments and applies differentiated learning. Our research question examines whether teachers differ with respect to their practice of formative assessment. As part of the project "Schools in Alpine Regions 2", we collected data from a voluntary sample of 280 teachers with multi-grade classes in two regions in Eastern Switzerland and one in Western Austria. We applied a mixed-method research design. A structural equation model reveals that teachers with multi-grade classes use a combination of formative assessment practices and tools in order to individualize and differentiate their teaching and instruction. The frequency with which formative assessment is practiced relates positively to the number of grades in a class and the teacher's self-concept regarding diagnostic observation competency. Results from two latent class analyses show two groups of assessment literacy, with only a small number using the full potential of formative assessment for learning. As part of the discussion, the results will be discussed with respect to their relevance for professional development and research.

Keywords Formative assessment; Mixed-Age teaching; Mixed-Methods; Rural schools

1 Einleitung

Obwohl Maier (2010) in seinem theoretischen Übersichtsartikel auf den Nutzen und die Bedeutung der *formativen Beurteilung* (formative assessment) für die Lernprozesse von Schülerinnen und Schüler aufmerksam gemacht hat, sind empirische Untersuchungen diesbezüglich in den deutschsprachigen Ländern weiterhin rar. Dies steht im Gegensatz zur breiten Rezeption der formativen Beurteilung auf internationaler Ebene, wo mit der Hattie-Studie (Hattie 2008) nochmals deutlich wurde, dass die formative Evaluation des Unterrichts und Feedback eine hohe Wirkung auf den Lernerfolg haben. In der Definition von Popham (2008, S. 6) finden sich die beiden Aspekte der Hattie-Studie zu einer Definition von formativer Beurteilung wie folgt zusammengefasst: Formative Beurteilung ist ein geplanter Prozess, bei dem Daten zu Schülertätigkeiten von der Lehrperson genutzt werden, um den Unterricht anzupassen oder um den Schülerinnen und Schülern zu ermöglichen, ihr Lernen zu steuern. Die Daten können sowohl auf schriftlichen Tests, wie auch auf bspw. Beobachtungen, Produkten, mündlichen Fragen oder Portfolios beruhen. Formative Beurteilung ist ein den Unterricht begleitender, laufender Prozess (Greenstein 2010). Als besonders hilfreich gilt die formative Beurteilung, wenn es darum geht, Lernen in heterogenen Gruppen zu unterstützen, bspw. in inklusiven oder jahrgangsgemischten Klassen (Hargreaves 2001; Klement 2005; Baeriswyl und Bertschy 2010; Beutel 2010).

In vielen kleineren, ländlichen Schulen findet sich aus Gründen der Schülerzahlen die Situation, dass in einem Klassenzimmer zwei oder mehr Jahrgänge gleichzeitig unterrichtet werden. Auch reformpädagogisch orientierte Schulen führen oft jahrgangsgemischte Klassen, aber mit pädagogischen Motiven (Wagener 2014; Ullrich 2015). Sind Kinder unterschiedlichen Alters gemeinsam in einer Klasse, führt dies zu einer grösseren Heterogenität als bei jahrgangsgetrennten Klassen. Damit steigt der Anspruch an eine Lehrperson, wenn es darum geht, den Stand des Lernprozesses der Kinder einzuschätzen, weil der soziale Vergleich innerhalb der Jahrgangsstufe erschwert wird. Als Folge, so Hargreaves (2001), rücken die individuellen Lernunterschiede verknüpft mit individuellen Lernzielen vermehrt in den Fokus der Lehrperson. Die individuellen Lernziele basieren wiederum auf der kriterialen Bezugsnorm, wie z. B. Lehrpläne und Bildungsstandards. Die Umsetzung einer formativen Beurteilung kann dabei verschiedene Formen und Werkzeuge beinhalten: z. B. Aufgaben zur Lernkontrolle am Ende einer Übungssequenz, Lehr-Lerngespräche in der Klasse, Beobachtung von Lernaktivitäten, Selbst- und Peerbeurteilung, Rubrics oder Portfolios. Dabei beschränkt sich die Tätigkeit der Lehrperson nicht nur auf geplantes Unterrichtshandeln, sondern auch spontane Beobachtungen und Rückmeldungen, die das Lernen der Schülerinnen und Schüler unterstützen, gehören dazu. Die so entstehenden Hinweise über die Lernfortschritte innerhalb einer Klasse ermöglichen es der Lehrperson, den Unterricht möglichst adaptiv (Houtveen et al. 1999) und differenzierend (Smit und Humpert 2012) zu gestalten. Formative Beurteilung dient somit nicht nur den Lernenden, sondern auch der Lehrperson und ist als Teil des didaktischen Arrangements von Lerngelegenheiten zu sehen (Perrenoud 1991).

Im Interreg-Projekt „Schulen im alpinen Raum 2“ wurden 280 Lehrpersonen aus kleineren Schulen mit jahrgangsgemischten Klassen aus der schweizerisch-österrei-

chischen Bodenseeregion mittels Fragebogen zu ihrer Unterrichtspraxis befragt. Teilweise amten die Lehrpersonen auch als Schulleitungsperson. Im Fragebogen gab es auch Items zur formativen Beurteilung. Anschliessend an die quantitative Erhebung fanden 75 vertiefende Interviews statt. Als zentrale Forschungsfragen galten dabei, ob Lehrpersonen Elemente einer formativen Beurteilung im jahrgangsgemischten Unterricht einsetzen und welchen Einfluss eine solche Beurteilung auf die Unterrichtsgestaltung hat. Weiter wurde untersucht, ob sich Gruppen von Lehrpersonen auf Grund ihrer Beurteilungskompetenz unterscheiden lassen. Gegebenenfalls lassen sich daraus verschiedene Entwicklungsstufen der formativen Beurteilung ableiten.

2 Theoretischer Hintergrund

2.1 Formative Beurteilung

Beurteilung und Didaktik sollen eine Einheit bilden (Perrenoud 1991). Während sich das Verständnis von Lernen seit der kognitiven Wende verändert hat und damit verbunden die didaktischen Arrangements vermehrt konstruktivistische Überlegungen miteinbeziehen, sind die Vorstellungen zur Nutzung von Beurteilung eher bei behavioristischen Vorstellungen stehen geblieben (Shepard 2000; Dochy 2001). Eine Didaktik, welche den Schüler als eigenständigen Lerner betrachtet (Reusser 1995), erfordert eine Beurteilung, welche den Lernprozess unterstützt und nicht nur den Lernstand überprüft. Es bestehen dabei enge Bezüge zu Lerntheorien wie z. B. Blooms „Mastery Learning“ (Guskey 2007) oder Vygotskys „Zone der nächsten Entwicklung“ (Black und Wiliam 2009). Ziel einer das Lernen unterstützenden Beurteilung ist es erstens, „Evidenzen“ zum momentanen Lernstand zu erzeugen und daraus mögliche didaktische Konsequenzen abzuleiten (Wiliam 2011). Zweitens soll dem Schüler in der Folge mittels Rückmeldung und didaktischem Arrangement erfolgreiches Lernen ermöglicht werden. Formative Beurteilung wird oft als zyklisches Modell gesehen (Roos 2001; Harlen 2007). Im Kreislauf-Modell von Cowie und Bell (1999) beginnt der Zyklus von formativer Beurteilung als Teil der Unterrichtsplanung mit dem Auswerten von Informationen zum Lernstand von Schülerinnen und Schülern.

Formative Beurteilung beinhaltet sowohl Fremd- wie die Selbststeuerung. Auch in der sozialkognitiven Lerntheorie von Bandura (1989) spielt externales Feedback eine bedeutsame Rolle bei der Steuerung des eigenen Lernens und damit verbunden auch bei der Entwicklung von Lernstrategien (vgl. Clark 2012). Hier findet sich auch ein Anschluss an die ansonsten eher pädagogischen Begründungen einer erweiterten, alternativen Beurteilung in deutschsprachigen Ländern: In von der Reformpädagogik bekannten offenen Lernformen werden Schülerinnen und Schüler als selbstständig Lernende betrachtet, die nicht nur Wissen, sondern auch sozial-kommunikative oder methodisch-strategische Kompetenzen erwerben. Solche erweiterten oder überfachlichen Kompetenzen, die bspw. bei Projekt- oder Freiarbeit eine bedeutsame Rolle spielen, wurden bislang eher selten in die Bewertung miteinbezogen und auch weniger rückgemeldet (Bohl 2004). Konkret soll auch der Lernprozess und damit verknüpft der individuelle Lernweg in die Beurteilung miteinbezogen werden (Winter 2004; Peschel 2005), womit wiederum das sozio-konstruktivistische Lernverständnis (Vygotsky 1974) angesprochen wird. Im konstruktivistisch orientierten Modell zur

Gestaltung von Unterricht von Dick und Carey aus den USA (Dick 1996) hat die formative Beurteilung auch einen prominenten Platz bei der Sicherstellung des Transfers des Lernens. Will die Lehrperson also bspw. sozial-kommunikative Kompetenzen fördern, bedarf es einer gezielten didaktischen Inszenierung, möglicherweise als Teil von Gruppenarbeit. Eingeplantes (Peer-)Feedback als Werkzeug der formativen Beurteilung kann dabei die Entwicklung entsprechender Kompetenzen fördern (Sadler 1989).

Formative Beurteilung und ihre Umsetzungsformen sind in allen Fächern wirksam (Black et al. 2004; Greenstein 2010). Zu unterscheiden sind bezüglich der Umsetzung weniger die Fächer als vielmehr die Komplexität der zu beurteilenden Kompetenz (McMillan 2010). Die formative Beurteilung schliesst sowohl fachdidaktische wie auch die oben erwähnten fachübergreifenden Aspekte mit ein. Fachdidaktisch, weil sie dem Schüler helfen soll, einen fachlichen Gegenstand zu erfassen, z. B. in den Naturwissenschaften (Keeley 2008), in der Mathematik (Besser et al. 2013) oder in der Sprache (Benjamin 2013). Um formative Beurteilung wirkungsvoll nutzen zu können, müssen Lehrpersonen über fach- und fachdidaktisches Wissen sowie über allgemeindidaktisches Wissen im Bereich der Beurteilung verfügen (Heritage 2007). Bei unvollständigem fach- und fachdidaktischem Wissen der Lehrpersonen kann es in Weiterbildungen sinnvoll sein, nur fachspezifische Beispiele formativer Beurteilung zu präsentieren (Schneider und Randel 2010).

Das oben erwähnte Wissen zur formativen Beurteilung wird benötigt, so Heritage (2007), um in der Klasse ein motivational günstiges Klassenklima herzustellen, welches es erlaubt, Selbst- und Peerbeurteilung zu nutzen, Fehler zu machen, und in welchem Unterschiede zwischen den Schülerinnen und Schülern akzeptiert sind. Zudem muss die Lehrperson in der Lage sein, die Schülerprodukte – mündliche, schriftliche, Demonstrationen – so zu nutzen, dass lernförderliche didaktische und pädagogische Unterrichtshandlungen sowie Rückmeldungen daraus abgeleitet werden können. Die Klärung des Schülerverstehens, der Fehlkonzepte, der Fertigkeiten und des Wissens geschieht dabei mit Bezug zu den intendierten Lernzielen und zu erwerbenden Kompetenzen. Grundlage der Schülerdaten sind dabei einerseits spontane, informelle Beobachtungen im Klassengespräch oder in individuellen Arbeitsphasen (Perrenoud 1991). Andererseits gehören dazu auch gezielte, formelle Beobachtungen während Präsentationsphasen oder Analysen von schriftlichen Arbeiten unmittelbar nach dem Unterricht.

Beim Wissen zur formativen Beurteilung im Bereich der Diagnose von bspw. Lernstand oder Fehlkonzepten gibt es Überschneidungspunkte zum Konzept der Pädagogischen Diagnostik. Auch bei der Pädagogischen Diagnostik sollen Voraussetzungen und Lernergebnisse ermittelt werden, um individuelles Lernen zu optimieren (Ingenkamp 2005, S. 13). Allerdings geht es hier auch um Förderungs-, Platzierungs- und Selektionsmassnahmen unter Verwendung wissenschaftlicher Methoden, was zumeist den Einsatz von Leistungstests oder Tests zur Klärung von sonderpädagogischem Förderbedarf beinhaltet. Die Konstruktion solcher Tests, der Einsatz und die entsprechende Diagnose sind allerdings in der Regel Aufgabe von Fachleuten, wie bspw. Sonderpädagoginnen und Sonderpädagogen oder Schulpsychologinnen und Schulpsychologen (vgl. Ricken 2007). Diese Fachleute können die Lehrpersonen sinnvoll unterstützen und eine vermehrte Zusammenarbeit beider ist wünschenswert,

insbesondere in inklusiven Klassen (Stein 2005). Während die Lehrperson Fachwissen und fachdidaktisches Wissen zur Beurteilung von Lernenden nutzt, kann eine Sonderpädagogin ihr Wissen zu Lernschwierigkeiten und Verhaltensproblemen für die Gestaltung weiterer Lernsituationen einbringen. Teilweise besteht auch die Notwendigkeit einer begrifflichen Klärung, wo diagnostische Förderung eingesetzt wird und wo es um didaktische, unterrichtliche Fragen geht. So ist Differenzierung des Unterrichts ursprünglich ein didaktisch-methodisches Konzept (Trautmann und Wischer 2009), welches aber in letzter Zeit auch als Fördermassnahme bezeichnet wird (Eckerth 2013).

Forschungsergebnisse zur formativen Beurteilung liegen analog zur theoretischen Rezeption hauptsächlich aus dem englischsprachigen Sprachraum vor. Nachfolgend werden einige, für die vorliegende Studie bedeutsame Ergebnisse zur Wirkung formativer Beurteilung und Feedback als Teilaspekt auf Leistung und Haltung präsentiert. In Ergänzung finden sich auch exemplarisch Ergebnisse zu Werkzeugen formativer Beurteilung: Rubrics und Portfolios. In der Metastudie von Black und Wiliam (1998b) zeigt sich, dass formative Beurteilungsstrategien einen Einfluss auf die Leistung haben. Die Umsetzungsqualität von formativer Beurteilung hat jedoch einen bedeutsamen Einfluss auf die Effektstärke (Yin et al. 2008). In einer weiteren Studie konnten Robinson, Myran, Strauss und Reed (2014) zeigen, wie eine wirksame Weiterbildung von Lehrpersonen in formativen Beurteilungsstrategien zu höherer Beurteilungskompetenz führt und damit verbunden zu höheren Schülerleistungen.

Die formative Beurteilung kann auch positiv auf die Schülerüberzeugungen wirken. Smit (2009) berechnete in einer Nachanalyse seiner Projekt-Daten mittels eines autoregressiven Pfadmodells, dass Lehrpersonen mit einer vermehrt formativ ausgerichteten Beurteilung einen positiven Effekt ($\beta = 0{,}43$) bezüglich Aspekten der Selbstkompetenz (Schüler-Skala mit Items zu Selbstvertrauen, Motivation und Nutzen der Rückmeldung) aufweisen; die aufgeklärte Varianz lag bei $R^2 = 0{,}19$ auf der Klassenebene. Ähnliches berichten Miller und Lavin (2007): In ihrer explorativen Längsschnitt-Studie zeigte sich eine positive Wirkung der formativen Beurteilung auf den Selbstwert und die Kompetenzeinschätzungen der Schülerinnen und Schüler. Feedback ist ein Teilaspekt der formativen Beurteilung. In der Metastudie von Hattie (2008) zeigt sich eine durchschnittliche Effektgrösse von 0,79 für Feedback auf die Schülerleistung. Allerdings ist nicht jede Form von Feedback gleich wirksam (Hattie und Timperley 2007). Die Wirksamkeit von Feedback hängt von der Wahrnehmung des Nutzers ab (Bangert-Drowns et al. 1991). Individuelles, schriftliches Feedback der Lehrpersonen führt zu höheren Mathematikleistungen als solches, bei dem die Rückmeldung mittels Note und damit rein summativ geschieht (Bürgermeister 2013; Harks et al. 2013). Hilfreiche Rückmeldungen lassen sich auch mit formativen Rubrics (Beurteilungsraster) erteilen. In einer Übersicht von Panadero und Jonsson (2013) finden sich sowohl Wirkungen von Rubrics auf die Leistungen
wie auch auf die Selbstwirksamkeits-Überzeugungen und die Selbstregulation.

Ein wichtiges Werkzeug von formativer Beurteilung sind Portfolios. Bezüglich der Effekte von Portfolios finden sich diverse Studien, bspw. eine Übersicht von Burner (2014), in der für das Fremdsprachen-Lernen gezeigt wird, dass Portfoli os sowohl bezüglich Motivation wie auch für die Schreibkompetenz wirksam sind. Bei Gläser-Zikuda und Lindacher (2007) ergab sich eine höhere Verwendung von Verarbeitungs-

und Monitoringstrategien in der Portfoliogruppe als in der Kontrollgruppe. Selbst- und Lehrer-Feedback kombiniert mit Portfolios erwies sich ebenfalls bei Baas et al. (2015) als hilfreich für die Nutzung von Strategien des selbstregulierten Lernens durch Schülerinnen und Schüler der Primarschule.

2.2 Jahrgangsübergreifendes/-gemischtes Lernen

In jahrgangs- oder altersgemischten Klassen sind Lehrpersonen während eines Schuljahres verantwortlich für den Unterricht von mehreren Alterskohorten (Little 2001). Gründe dafür können organisatorischer Art sein, wenn bspw. die Schülerzahlen klein sind oder sinken. Es können aber auch pädagogische Überlegungen sein, die zu einem jahrgangsgemischtem oder -übergreifendem Lernen führen: Die Vielfalt der altersgemischten Schülerinnen und Schüler wird für das Lernen als Chance gesehen. Pädagogische Vorteile liegen im sozialen und demokratischen Lernen (vgl. Achermann und Gehrig 2011; Wagener 2014). Die didaktische Umsetzung von Unterricht, in welchem unterschiedliche Jahrgänge miteinander am gleichen Thema arbeiten, wird im englischen Sprachgebrauch als mixed-age oder multi-age teaching bezeichnet (Katz et al. 1990; Cornish 2006). Ein Pluspunkt eines solchen Unterrichts wird in der Möglichkeit gesehen, dass Kinder dem Lerngegenstand mehrmalig innerhalb eines mehrjährigen Zyklus begegnen (Spiralprinzip), wobei dann immer der jeweilige Entwicklungsstand der Schülerinnen und Schüler zu berücksichtigen ist (Bruner 1960). Im jahrgangsübergreifenden Unterricht soll kindzentriert und kooperativ gelernt werden, das Lernen soll individuell dokumentiert werden und eine intensive Beratung durch die Lehrperson soll erfolgen. Unterrichten in jahrgangsgemischten Klassen ist anspruchsvoll und vielen Lehrpersonen fehlt die Erfahrung für einen entsprechenden Unterricht (Miller 1991; Little 2001). So gibt es viele Lehrpersonen, die insbesondere die Kernfächer (Sprache und Rechnen) weiterhin grundsätzlich innerhalb der gemischten Klasse in jahrgangsgetrennten Gruppen unterrichten (Raggl 2011). Es finden sich aber auch Lehrpersonen, welche die Altersmischung didaktisch nutzen. Dazu gibt es fachspezifische Überlegun gen, z. B. in der Mathematik bei Nührenbörger and Steinbring (2009) oder in der Sprache bei Marley et al. (2011). Die meisten empirischen Untersuchungen zu jahrgangsgemischten Klassen finden sich auf den unteren Schulstufen. Viele Ergebnisse sind neutraler Art, sie zeigen keine oder nur geringe positive Effekte der Jahrgangsmischung auf die Schulleistung oder bezüglich sozio-emotionaler Merkmale (Veenmann 1995; Kuhl et al. 2013). Allerdings finden sich auch keine allenfalls befürchteten negativen Effekte. Hattie (2002) merkt zu den geringen Effektgrössen kritisch an, dass die wenigsten Lehrpersonen bei der Umstellung auf die Jahrgangsmischung ihren Unterrichtsstil ändern. Zudem wurde auch die didaktische Inszenierung des jahrgangsgemischten Lernens bis jetzt nie in Studien zur Wirkung miteinbezogen (Kuhl et al. 2013).

2.2.1 Formative Beurteilung im jahrgangsgemischten Unterricht

Schon in Klassen mit einem Jahrgang ist die Lehrperson gefordert, wenn sie sich um alle Kinder und ihre individuellen Lernprozesse kümmern will. In der altersgemischten Klasse kommt nun noch dazu, dass die Kinder auch entwicklungsmässig und vom

Wissen her unterschiedliche Ansprüche an die Lehrperson stellen (Lang et al. 2010). Zudem ist in einer Klasse mit verschiedenen Jahrgängen die Spannbreite der Lernziele breiter. Oftmals reicht es innerhalb einer Lektion zeitlich nicht, während den Arbeitsphasen bei allen Kindern vorbeizugehen und deren Probleme und Lernzuwachs wahrzunehmen. Die Lehrperson kann jedoch einen Gesamteindruck über den Fortschritt der Klasse respektive über den der Jahrgangs- oder Niveaugruppen gewinnen. Will die Lehrperson Lerngruppen auf Grund des Leistungsstands bilden, bedarf es wiederum an durch formative Beurteilung gewonnener Informationen (Eckerth 2013). Zusätzliche Hinweise bekommt die Lehrperson, wenn sie die Aufgaben im Schülerheft oder den Fortschritt auf dem Wochenplan kontrolliert. Will die Lehrperson jedoch bestimmte Kompetenzen bei allen Kindern prüfen, muss sie Beurteilungssituationen planen, dann systematisch Dokumente sammeln und Handlungen beobachten (Beutel 2010). Dies sollte so geschehen, dass die Informationen später auch einfach für Aussagen zum Lernen des Kindes nutzbar sind und nicht noch viel Zeit für die Aufbereitung der Informationen benötigt wird. Ansonsten kann die Leistungsdokumentation – bspw. in einem Portfolio (Stone 1996) – auch zu einer zeitlichen Überforderung der Lehrperson führen. Entsprechend bedürfen Lehrpersonen im jahrgangsgemischten Unterricht guter Kompetenzen im Bereich der formativen Beurteilung. Da die Schülerinnen und Schüler häufiger selbstreguliert arbeiten, benötigen auch die Lernenden vermehrt Beurteilungs-Kompetenzen bei der Selbst- und Peerbeurteilung (Little 2007; Beutel 2010). Mulryan-Kyne (2007) nennt für die Lehrerbildung im Bereich der Beurteilung in jahrgangsübergreifenden Klassen folgende Ausbildungsziele: Schülerlernen beobachten und beurteilen, ein wirksames Konzept für eine den Lernprozess begleitende Beurteilung entwickeln, lernförderliche und gezielte Rückmeldungen zum Lernen geben, die Ergebnisse von Lernkontrollen für das Lernen der Schülerinnen und Schüler nutzen sowie die individuelle Entwicklung unter Verwendung von lernförderlichen Kommentaren dokumentieren können. Aus den Vorschlägen von Mulryan-Kyne (2007) lässt sich zusammenfassend ein Konstrukt der formativen Beurteilung im jahrgangsübergreifenden Unterricht ableiten und zur Analyse auf die Forschungsdaten anwenden:

- Formative Beurteilung zeigt sich im Unterricht in der Verwendung gezielter, auf Kriterien beruhender Beobachtung und der Dokumentation und Beurteilung dieser Beobachtungen.
- Die Daten werden genutzt zur Unterrichtsplanung und zur individuellen Rückmeldung.
- Zur Dokumentation und Rückmeldung werden Portfolios genutzt.
- Die Schüler üben und nutzen Selbstbeurteilung.

2.3 Forschungsfragen

1. Zeigen sich Elemente der formativen Beurteilung als Teil eines Gesamtkonstrukts im jahrgangsgemischten Unterricht der Primarstufe?
2. Gibt es einen Einfluss der formativen Beurteilung auf die Unterrichtsplanung (inkl. Differenzierung/Individualisierung und Selbstkompetenz fördern)?
3. Lassen sich Gruppen von Lehrpersonen bezüglich ihrer formativen Beurteilungskompetenz identifizieren?

3 Methoden

3.1 Stichprobe

Für die Stichprobe wurden alle Lehrpersonen mit Mehrjahrgangsklassen der Grundstufe/Primarstufe in Vorarlberg, Graubünden und St. Gallen angefragt; zusätzlich dazu auch die jeweiligen Schulleiter/-innen. Während die Grundstufe in Vorarlberg die 1. bis 4. Jahrgangsstufe beinhaltet, umfasst die Primarstufe in den Schweizer Kantonen zudem noch die 5. und 6. Klasse. Die Beteiligung lag für Vorarlberg bei rund 50 % und für die beiden Schweizer Kantone bei rund 30 % der angefragten Lehrpersonen. Die Stichprobengrösse beträgt 280 Personen (207 Lehrpersonen und 73 Schulleiter/-innen). Teilweise haben die Befragten eine Doppelfunktion inne. Der überwiegende Teil der Lehrpersonen unterrichtet in kleinen, ländlichen Schulen, da dort die Schülerbestände oft das Zusammenlegen von Jahrgängen erfordern. 73 % der Befragten sind weiblichen Geschlechts und das mittlere Alter beträgt 42 Jahren. Die Dienstzeit liegt zwischen einem halben Jahr und 43 Jahren, mit einem Durchschnitt von 18 Jahren. Die Hälfte der teilnehmenden Schulen weist eine Schülerzahl von unter 50 aus, etwa 10 % der Schulen hat mehr als 100 Schülerinnen und Schüler. Obwohl diese 10 % nicht die typischen kleinen, ländlichen Schulen sind, wurden sie aufgrund der ausgeprägten jahrgangsübergreifenden Umsetzung in die Stichprobe integriert. Aus dieser Stichprobe und ergänzend auf der Basis persönlicher Empfehlungen von u. a. Personen aus der Aus- und Weiterbildung wurden 30 Schulleiterinnen und Schulleiter und 45 Lehrpersonen aus allen drei Regionen (je 10 Schulen) für die vertiefenden Interviews ausgewählt. Aus Sicht des Mixed-Methods-Vorgehens wurde zuerst eine grössere quantitative und anschliessend gezielt eine qualitative Stichprobe erhoben (Creswell et al. 2011, S. 179).

3.2 Untersuchungsdesign und Instrumente

Das zweite Forschungsprojekt zu „Schule im alpinen Raum“ dauerte von Anfang Juni 2012 bis Ende Mai 2015. Methodisch hatte das Projekt ein sequentielles, erläuterndes Mixed-Methods-Design (Creswell und Plano Clark 2011, S. 185) mit einer einmaligen quantitativen Befragung, auf welcher aufbauend eine qualitative Interviewbefragung folgte. Als Rahmen für die Erhebung diente ein Input-Output Modell (Scheerens 2000). Zum Input gehören z. B. Schulgrösse; materielle, räumliche und personelle Ressourcen; Arbeitszufriedenheit usw. Prozessvariablen betrafen die Schule (Führung, Teamkooperation usw.) und die Klasse (Unterrichtsgestaltung, Klassenklima usw.). Outputvariablen wurden nur in den Interviews erhoben (Schülerüberzeugungen).

3.2.1 Fragebogen

Es wurde ein Fragebogen für die Lehrpersonen und Schulleitungspersonen auf der Basis der klassischen Testtheorie entwickelt, überprüft und einmalig eingesetzt. Der Fragebogen enthielt etwa 200 Items bezüglich Person, Kontext, Schulentwicklung

und Unterrichtsgestaltung. Die meisten Items wurden zusammengefasst in 35 Skalen; alle mit einem Cronbach alpha > 0,70. Zusätzlich enthielt der Fragebogen Einzelitems und einige offene Fragen, insbesondere für Personen mit zusätzlichen Schulleitungsaufgaben. Die Items stammten zu einem grossen Teil aus der ersten Untersuchung zu „Schule im alpinen Raum“ (Müller et al. 2011). Items zur formativen Beurteilung konnten teilweise aus dem Dissertationsprojekt von Smit (2009) übernommen werden. Das Antwortformat der Likert-Skalen war zumeist fünf oder sechsstufig (absolute Zustimmung = 5/6, absolute Ablehnung = 1). In Tab. 1 sind die deskriptiven Angaben der Skalen zur Beschreibung des jahrgangsgemischten Unterrichts aufgeführt. Zusätzlich wurde die Kompetenzeinschätzung der Lehrperson zum Umgang mit Heterogenität erfragt. Als Kontrollvariablen bei den Lehrpersonen wurden Jahrgangsstufe(n), Geschlecht, Pensum und Dienstalter verwendet.

Die Itemformulierungen, welche für die Klärung der Forschungsfragen notwendig sind, finden sich direkt im Zusatzmaterial.

Tab. 1 LP-Skalen zur jahrgangsgemischten Unterrichtsgestaltung mit den Produkt-Moment-Korrelationen

	M	SD	1	2	3	4	5	6	7	8	9
1 Kompetenz Heterogenität[a]	4,65	0,63	–	–	–	–	–	–	–	–	–
2 Rolle als Lerncoach[a]	5,06	0,63	0,38	–	–	–	–	–	–	–	–
3 Individualisierung	3,72	0,86	0,47	0,31	–	–	–	–	–	–	–
4 Lernen am gemeinsamen Thema/Inhalt	3,67	0,74	0,22	0,17	0,19	–	–	–	–	–	–
5 Flexible Lerngruppen	3,37	0,65	0,24	0,28	0,30	0,22	–	–	–	–	–
6 Sozialkompetenz Klasse[a]	4,55	0,62	0,51	0,24	0,22	0,23	0,13	–	–	–	–
7 Methodenkompetenz Klasse[a]	4,30	0,65	0,41	0,13	0,20	0,19	0,09	0,69	–	–	–
8 Beobachtung/ Diagnose	3,62	0,67	0,54	0,53	0,49	0,24	0,34	0,36	0,27	–	–
9 Schüler Selbstbeurteilung	3,38	0,80	0,36	0,36	0,36	0,11	0,20	0,27	0,24	0,38	–
10 Beurteilungsdokumentation	3,16	0,86	0,42	0,37	0,51	0,17	0,21	0,26	0,25	0,58	0,47

Anmerkung: $n = 254$

[a]Sechsstufig, sonst fünfstufige Likertskala, hohe Mittelwerte bedeuten hohe Zustimmung oder grosse Häufigkeit, alle Cronbach alpha > 0,70; alle $r > 0{,}16$ sign. mit $p < 0{,}01$

3.2.2 Interview

Die vertiefenden Leitfadeninterviews umfassten fünf Themenbereiche: 1. Einstieg (Erleben, Schule), 2. Jahrgangsgemischter Unterricht – Gestaltung des Unterrichts, 3. Rahmenbedingungen, Unterstützung, 4. Rolle der Schulleitung in einer kleinen

Schule, 5. Fragen zur Person und zur Berufsbiographie. Beurteilung war ein Teilaspekt des Themas Unterrichtsgestaltung (siehe Kategoriensystem im Zusatzmaterial). Zusätzlich gab es noch Interviews mit Schülergruppen zu den Themen Helfen, Wohlbefinden und Wünsche. Zeitlich dauerten die Interviews zwischen 45 und 60 min. In der Regel fanden die Interviews anschliessend an einen Unterrichtsbesuch statt. Damit konnten auch Rückfragen zu Beobachtungen der interviewenden Person mit in die Befragung einfliessen.

3.3 Analysen

Zur Klärung der Zusammenhänge zwischen den latenten Variablen wurden Strukturgleichungsmodelle (SEMs) unter Verwendung der Software M-Plus 7 (Muthén und Muthén 2012) berechnet. Diese Modelle stellen eine Kombination von Faktoranalyse und Pfadmodellen dar und eignen sich zur Prüfung theoretischer Modelle. M-Plus verwendet standardmässig das FIML Schätzverfahren zur Einbeziehung fehlender Werte. Der Modell-Fit wurde mit Hilfe verschiedener goodness-of-fit Kennwerte bestimmt: Chi Quadrat, Bentlers komparativer fit Index (CFI), der Tucker-Lewis Index (TLI) und der statistische Mittelwert der Standardfehler (RMSEA). CFI und TLI Werte von 0,95 und darüber und RMSEA Werte von 0,05 und darunter weisen auf einen guten Modell-Fit hin (Schermelleh-Engel et al. 2003). Die transkribierten Interviews wurden mit Hilfe der qualitativen Inhaltsanalyse (Mayring 2000) kategorisiert und verdichtet. Dabei wurden deduktive und induktive Codes verwendet. Verständnisfragen wurden im Team bearbeitet, die Kodierung fand einzeln statt. Zum Einsatz kam für diesen Zweck die Software MAXQDA (VERBI Software. Consult. Sozialforschung 2012). Sowohl für die dichotomen Daten (kommt vor/kommt nicht vor) aus den Interviews wie auch für die likert-skalierten Fragebogendaten wur den mit M-Plus Latent-Class-Analysen (LCA) zur Klassenbildung durchgeführt. Mit LCA können Antwortmuster von Subgruppen bestimmt werden (Geiser 2010). Werden LCA mit stetigen Daten gerechnet, bezeichnet man eine solche LCA auch als „Latent Profile Analysis“ (LPA). Der Unterschied zeigt sich u. a. in der Angabe der geschätzten Variablenwerte für jede Klasse. Bei der LCA mit dichotomen Daten wird der Wert pro Variable und Klasse in Wahrscheinlichkeiten ausgedrückt. In der LCA mit stetigen Daten (= LPA) ergibt sich für jede Variable ein geschätzter Mittelwert pro Klasse. Die qualitative Auswertung wird hier aus Platzgründen nur beschränkt dargestellt. Sie dient im Sinne eines Mixed-Method-Ansatzes (Bazeley 2009) der Synthese gemeinsamer Datenquellen. Dieses Vorgehen erlaubt es, den Forschungsgegenstand ganzheitlicher zu untersuchen.

4 Ergebnisse

4.1 Ergebnisse aus den Fragebogendaten

4.1.1 Strukturgleichungsmodell zur formativen Beurteilung

Rund 40 % der Lehrpersonen in unserer Fragebogen-Stichprobe arbeiteten mit zwei Jahrgängen in einer Klasse. Weitere 25 % hatten drei oder vier Jahrgänge in einer

Klasse, während 5 % sogar mit fünf Jahrgängen in einer Klasse arbeitete. Der restliche Teil der Befragten arbeitete als Fachlehrpersonen in verschiedenen Klassen. Nur 12 % gaben an, dass sie nie oder selten zu einem gleichen Thema jahrgangsgemischten Unterricht durchführten; 24 % nutzten jahrgangsgemischte Unterrichtssettings gelegentlich. Der Rest gab an, häufig oder immer jahrgangsgemischten Unterricht durchzuführen. Die Lehrpersonen schätzen Items zur eigenen Kompetenz der Unterrichtsgestaltung in heterogenen Klassen als eher hoch ein (siehe Tab. 1). Dazu gehört bspw. die Diagnose des Lernstandes, die Gestaltung differenzierender Aufgaben oder das Aufbereiten eines Themas für verschiedene Jahrgänge.

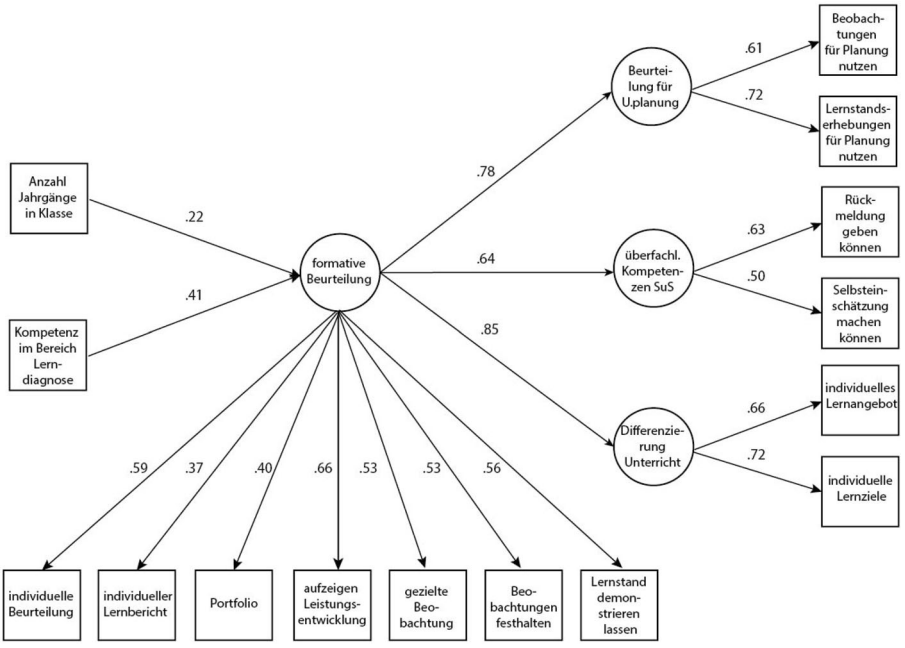

Abb. 1 SEM zu formativer Beurteilung und Aspekte der Unterrichtsplanung (alle Pfade sind signifikant, $p < 0{,}05$; standardisierte Schätzwerte)

Zur Klärung des zu messenden Konstrukts „formative Beurteilung“ wurden mit den Items der Skalen „Beobachtung/Diagnose“, „Schüler Selbstbeurteilung“ und „förderorientierte Beurteilung“ theoriegeleitet verschiedene konfirmatorische Faktorenanalysen (CFA) gerechnet. Die latente Variable „formative Beurteilung“ besteht beim besten Modell aus den vier Items der Skala „Beurteilungsdokumentation“ sowie drei Items aus der Skala „Beobachtung/Diagnose“. In Abb. 1 sind die Items inhaltlich ersichtlich. Die Fit-Werte für das CFA-Modell „formative Beurteilung“ sind gut: $X^2 = 23{,}29$, df = 14, $p = 0{,}06$, CFI = 0,97, TLI = 0,95 und RMSEA = 0,05.
Für das Modell konnten 40 Personen der Stichprobe von $N = 280$ nur als Schätzwerte (FIML-Verfahren) berücksichtigt werden, da fehlende Werte auf allen oder den x-Variablen vorlagen. Als Nächstes wurde geprüft, welche Kontextvariablen einen Einfluss auf die Häufigkeit des Einsatzes einer formativen Beurteilung haben. Es zeigt

sich, dass die Selbsteinschätzung der eigenen Kompetenz zur Lerndiagnose und etwas schwächer auch die Anzahl der Jahrgänge in einem positiven Zusammenhang mit der Häufigkeit der formativen Beurteilung steht. Andere Kontextvariablen wie Geschlecht, Dienstalter usw. waren dagegen nicht signifikant. Gemäss theoretischen Modellen für die Gestaltung des Unterrichts, z. B. Dick (1996), ist die formative Beurteilung ein wichtiger Bestandteil der Unterrichtsplanung.

Die formative Beurteilung wird oft als zyklisches Modell gesehen (Roos 2001; Harlen 2007). Im Modell von Cowie and Bell (1999) beginnt der Zyklus von formativer Beurteilung als Teil der Unterrichtsplanung mit dem Auswerten von Informationen zum Lernstand von Schülerinnen und Schülern. In diesem Sinne wählen wir die Richtung der Pfade im Modell ebenfalls ausgehend von der formativen Beurteilung hin zur Unterrichtsplanung. Dabei ist die Richtung in einem SEM nicht im Sinne einer generellen Ursachenerläuterung, sondern ähnlich zu einem regressiven Zusammenhang einer Variablen mit einer andern Variablen zu verstehen (Bollen und Pearl 2013, S. 12). Die Wirkungsrichtung im Modell ist jedoch bedeutsam und dient als Teil der wissenschaftlichen Erkenntnisgewinnung. Nach Stefanikis Harris (2011) sind auch Differenzierungsmassnahmen abzuleiten von formativer Beurteilung. Zudem wird angenommen, dass die formative Beurteilung einen Einfluss auf die Lernkompetenz resp. das selbstregulierte Lernen der Schülerinnen und Schüler hat (Nicol und Macfarlane-Dick 2006). Aus diesem Grund wurde das CFA-Modell um weitere Variablen resp. Items aus den Skalen zum Unterricht (siehe Tab. 1) zu einem SEM-Modell erweitert (siehe Abb. 1) und in verschiedenen Ausführungen geprüft. Das finale Modell weist die folgenden Fit-Werte auf: $X^2 = 125{,}17$, df = 83, $p = 0{,}00$, CFI = 0,94, TLI = 0,92 und RMSEA = 0,05 und besitzt damit leicht schlechtere Werte als das CFA-Modell. Allerdings vermag es mehr Zusammenhänge zu erklären. So zeigt sich im finalen Modell, dass Lehrpersonen, welche häufiger eine formative Beurteilung nutzen, diese auch häufiger für die Unterrichtsplanung und die Differenzierung verwenden. Zudem weisen die Lernenden nach Angaben der Lehrpersonen mit häufigerer Nutzung von formativer Beurteilung auch höhere Selbstkompetenzen auf. Gesamthaft erklärt das Modell rund 23 % der Varianz bezüglich der latenten Variable „formative Beurteilung“, 42 % der Varianz bezüglich der Einschätzung der beiden überfachlichen Kompetenzen und 72 % bezüglich der Differenzierung respektive Individualisierung.

4.1.2 Latent Profile Analyse basierend auf den Fragebogendaten

Mit den Items der Fragebogen konnten 7 Skalen zur formativen Beurteilung gebildet werden, welche auf einer fünfstufigen Skala von den Lehrpersonen eingeschätzt wurden (siehe Abb. 1). Auf der Basis einer Latent Class Analyse (vgl. Vermunt und Magidson 2002) liess sich klären, ob sich anhand dieser Merkmale Gruppen von Lehrpersonen unterscheiden lassen, bspw. eine solche mit traditioneller und eine mit erweiterter Beurteilung (vgl. Grunder und Bohl 2001). LCA mit kontinuierlichen Daten werden auch Latent Profile Analysen (LPA) genannt. In beiden Beurteilungsformen zeigen sich die Elemente formativer Beurteilung, aber die Gruppe mit erweiterter Beurteilung setzt die formative Beurteilung häufiger ein. Wir verglichen dazu eine Va-

riante mit zwei Gruppen mit einer Variante mit drei Gruppen. Das Zwei-Gruppen-Modell erbrachte minimal tiefere Fit-Werte mit einem BIC von 5227,92, einem AIC von 5146,02 und einem aBIC von 5155,00. Das Drei-Gruppen Modell wies ähnliche Werte auf: BIC = 5235,61, AIC = 5146,59 und aBIC 5156,35. Da in einer zweiten LPA auf der Basis von Interviewdaten ein Zwei-Gruppen-Modell erstellt wurde (siehe weiter unten), wird aus Gründen der Synthese der Ergebnisse im Folgenden auf das Zwei-Gruppen-Modell abgestützt. In der Abb. 2 sind die Skalen samt den durchschnittlichen Mittelwerten der jeweiligen Gruppe abgebildet. Beiden folgenden drei Codes zeigt sich für die Gruppe mit erweiterter Beurteilung ein Wert im Skalenbereich „oft", während die Gruppe mit traditioneller Beurteilung einen Wert im Bereich zwischen „selten" und „manchmal" zeigt: 1. Lernentwicklung aufzeigen, 2. schriftliches Festhalten von Beobachtungen, 3. Individuelle Beurteilung. Portfolios werden von beiden Gruppen selten eingesetzt.

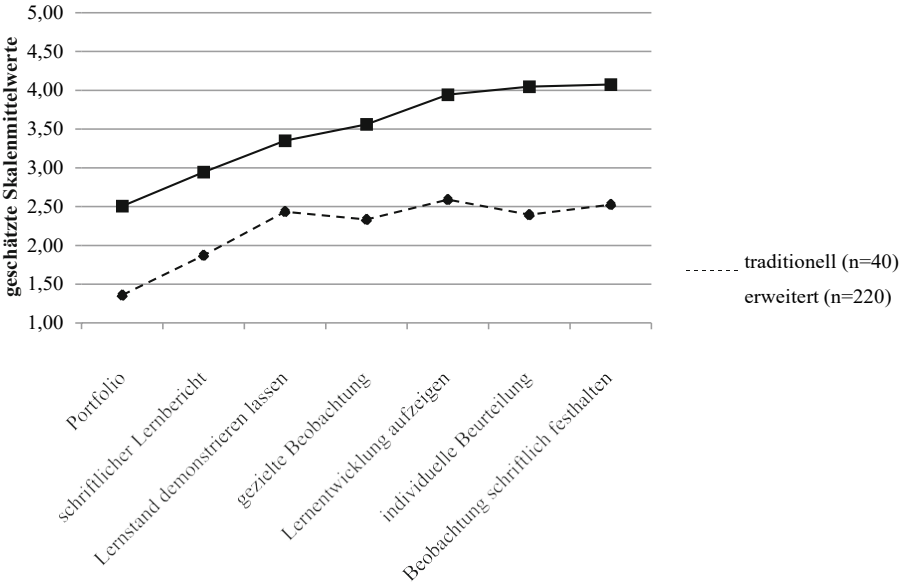

Abb. 2 Wahrscheinlichkeiten der Gruppenzugehörigkeit für ein 2-Klassen LCA-Modell basierend auf Fragebogenskalen zu Aspekten der formativen Beurteilung. Die Antwortkategorien lauten: 1 (nie), 2 (selten), 3 (manchmal), 4 (oft), 5 (immer)

4.2 Ergebnisse aus den Interviewdaten

In den Interviews mit den Lehrpersonen finden sich konkrete Beispiele, wie die Lehrpersonen die formative Beurteilung nutzen, um den Unterricht für einzelne Schülerinnen und Schüler resp. Schülergruppen zu planen. Diese Beispiele sind teilweise Kategorien, die anschliessend in die Latent Class Analyse Eingang finden. Im ersten Beispiel (GR10) wird mittels formativer Lernkontrolle das Vorwissen geklärt. Darauf aufbauend werden dann die nächsten Übungsaufgaben ausgewählt. Im zweiten Beispiel (V1) nutzt die Lehrperson eine formative Lernkontrolle, um nach einer gewissen Übungszeit die nächsten Lernziele zu bestimmen. Dazu lässt sie sich vom Kind die Bearbeitung einer Aufgabe vordemonstrieren.

> Und danach mache ich auch, zum Beispiel bei Kilogramm oder bei Nomen, wie einen Einstiegstest, um zu sehen, ob die das schon können. Die müssen dann gar nicht gross an dem Thema arbeiten. Schon am Thema, aber auf ganz
>
> anderem Niveau. Also ich habe zwei, die wirklich sehr stark sind in Deutsch, die können eigentlich den Stoff der 4. Klasse schon lange. Und dann teste ich am Anfang einfach schnell bei allen, und sehe auch gleich, wo sie noch Mühe haben und wie viel sie noch üben müssen. Das ist auch für die Kinder gut. Sie sehen, was sie schon können und wo sie noch weiter arbeiten müssen (GR10/ Vorwissen klären).
>
> ..., die Kinder kommen, wenn sie eine bestimmte Arbeit geübt haben, zu mir und wollen eine Lernzielkontrolle machen. Oder ich sage: „Mensch, ich glaube, du hast schon so viel geübt, komm, jetzt probieren wir das gerade miteinander." Ich schreibe mir dann jeweils von jedem Kind auf, wann es welche Lernziele erreicht hat (V1/formative Lernkontrolle).

Erfahrung spielt eine wichtige Rolle bei der Beurteilung der Lernenden und ihrer Entwicklung. Doch was heisst eigentlich „Erfahrung"? Das, was die Lehrpersonen äussern, wenn sie Erfahrung meinen, ist eigentlich implizites Wissen. Sie können zwar einschätzen, was der Leistungsstand eines Schülers oder einer Schülerin ist, aber sie können das nicht explizit belegen. Die Lehrpersonen berichten in der Folge von subjektiven Eindrücken und zufälligen Beobachtungen.

> Ja, sagen wir beim Lesen lernen. Wenn man schon so viele Kinder beobachtet hat, dann hat man vielleicht auch eine gewisse Gelassenheit und weiss, irgendwann kommt dann das schon. Man kann vielleicht einschätzen, wie so ein Lernprozess geht, auch von der Zeit her. Das gibt schon eine gewisse Erfahrung und auch das Vertrauen, dass man weiss, es ist auf dem richtigen Weg, es kommt schon noch rechtzeitig (GR7b/Erfahrung).

Problematisch ist, dass bei der Einschätzung der Leistung aus der Erfahrung leicht Beurteilungsfehler entstehen können. Insbesondere Voreingenommenheiten, wie bspw. globale Bewertungen einer Schülerin oder eines Schülers über alle Fächer hinweg (Halo-Effekt) sind häufiger als Lehrpersonen meinen (Ingenkamp 1995). Die Erfahrung wird oft ergänzt durch Vergleiche mit andern Schülerinnen und Schülern in

der Klasse. Dabei wird der Vergleich mit Kindern aus andern Jahrgängen als vorteilhaft empfunden. So meint eine Lehrperson:

> Ja, einerseits ist es leichter als in einer Regelklasse, denn man hat den direkten Vergleich mit den älteren Schülern. Man sieht, ob einer (...), ich kann sagen, dieser Schüler ist ein Viertklässler, aber er ist schon reif, er könnte schon etwas mit den Fünftklässlern machen. Das ist einfacher, wenn man die Unterschiede sieht. Zudem muss selbstverständlich die Lehrperson für die Beobachtung geschult sein, dies gilt aber auch in der Regelklasse, sie muss alle Besonderheiten notieren, die es gibt (GR4/soziale Bezugsnorm).

Alle Lehrpersonen beobachten ihre Schülerinnen und Schüler während Arbeitsphasen. Die einen nehmen die Beobachtungen zur Kenntnis, die andern verschriftlichen sie. Letzteres kann sich auf bestimmte ausgewählte Situationen beschränken oder generell gemacht werden. Lehrpersonen können ihre Beurteilungsqualität steigern, wenn sie ihre Urteile mit denen anderer Lehrpersonen überprüfen. Dies kann innerhalb des Schulhauses, bspw. mit der Heilpädagogin, aber auch in Zusammenarbeit mit Lehrpersonen anderer Schulhäuser geschehen. Im folgenden Abschnitt berichtet eine Lehrperson einer 6. Klasse aus der Schweiz, wie sie die Zusammenarbeit, hier im Hinblick auf den Übertritt in die Sekundarstufe, gestaltet:

> Ich mache mit Z. (Anm.: mit dem grösseren Nachbardorf) Gespräche wegen des Übertritts, die (Anm.: Schüler/-innen) kommen ja nachher zusammen und wir haben das abgemacht, dass die Lehrer von der sechsten Klasse in Z. und ich zusammenarbeiten. Zwei, drei Repetitionen machen wir gemeinsam und dann vergleichen wir (GR11a/Austausch mit LPs).

4.2.1 Latent Class Analyse basierend auf den Interviewdaten

In der qualitativen Inhaltsanalyse konnten 13 Codes für Aspekte einer formativen Beurteilung, die mehrmals (zwischen 3 und 29 mal) genannt wurden, identifiziert werden. Mit Hilfe einer Latent Class Analyse (vgl. Vermunt und Magidson 2002) liess sich klären, ob sich anhand dieser Merkmale Gruppen von Lehrpersonen unterscheiden lassen, bspw. eine solche mit traditioneller und eine mit erweiterter Beurteilung (vgl. Grunder und Bohl 2001). In beiden Beurteilungsformen zeigen sich Elemente formativer und summativer Beurteilung, aber die Gruppe mit erweiterter Beurteilung setzt die formative Beurteilung systematischer, zielorientierter, vielseitiger und konsequenter ein. Wir verglichen dazu eine LCA-Variante mit zwei Gruppen mit einer mit drei Gruppen (n = 75). Das Zwei-Gruppen-Modell erbrachte leicht höhere Fit-Werte mit einem BIC von 589,40, einem AIC von 656,61 und einem aBIC von 565,21. Das Drei-Gruppen Modell wies die etwas besseren, tieferen Werte BIC = 591,18, AIC = 693,15 und aBIC 554,47 auf. Da die dritte Gruppe bei der Drei-Gruppen-Variante nur aus 3 Personen besteht, wird der Zwei-Gruppen-Variante den Vorzug gegeben. Bei der Drei-Gruppen-Variante finden sich besonders für die dritte Gruppe auch eher viele Boundary Estimates und geringe mittlere Klassenzuordnungswahrscheinlichkeiten, was auf Schwierigkeiten mit der Zuordnung hindeutet (Geiser 2010). In der Abb. 3 sind die Codes samt der Wahrscheinlichkeit, mit der sie

bei einer der zwei Gruppen auftreten, abgebildet. Bei vier Codes zeigt sich ein deutlicher Unterschied: 1. Formative Lernkontrollen, 2. schriftliches Festhalten von Beobachtungen, 3. Individuelle Beurteilung und 4. Planung des Unterrichts unter Berücksichtigung von Erkenntnissen aus Beurteilungen. Für diese Merkmale besteht eine hohe Wahrscheinlichkeit, dass sie der Gruppe „erweiterte Beurteilung“ zugehören. Die vier Merkmale werden von den Lehrpersonen dieser Gruppe häufig verwendet, von der traditionellen nicht. Andere Merkmale wie die Peer-Beurteilung oder das Portfolio werden von beiden Gruppen selten eingesetzt. Zu den vier häufigen Merkmalen der Gruppe „erweiterte Beurteilung“ gehört auch das oben erwähnte Festhalten von Beobachtungen. Im folgenden Interviewausschnitt berichtet eine Lehrperson, dass die Dokumentation der Beobachtungen bedeutsam ist für das Erfassen von Entwicklungen. Interessant ist auch der Hinweis zum Bildungsstandard und dass sich mit dem Einsatz eines nationalen Leistungstests nichts über die individuelle Entwicklung aussagen liesse.

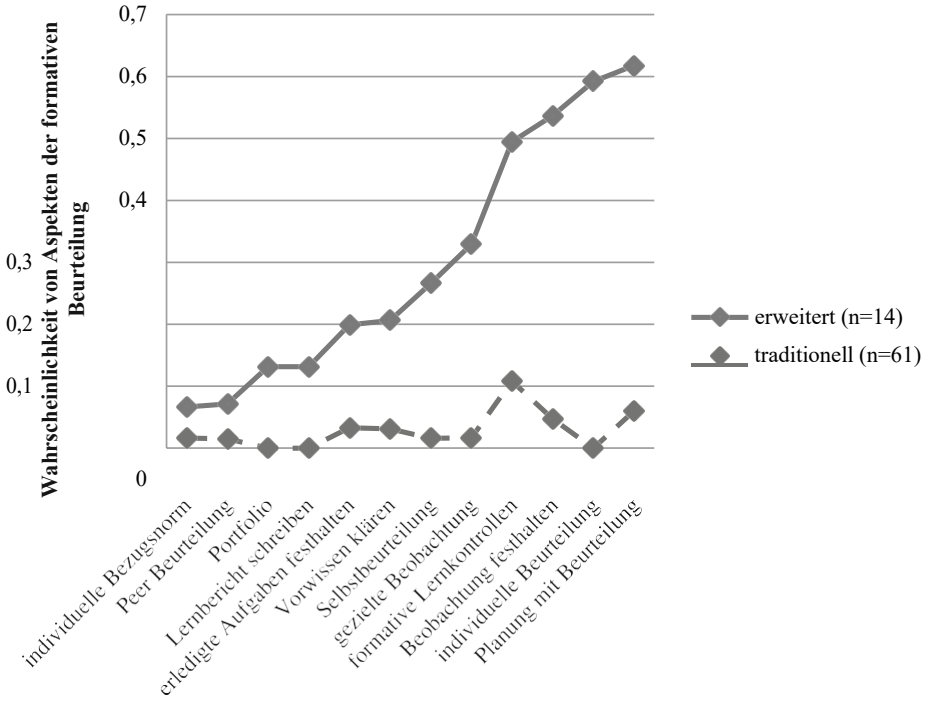

Abb. 3 Wahrscheinlichkeiten der Gruppenzugehörigkeit für ein 2-Klassen LCA-Modell basierend auf Interviewkodierungen zu Aspekten der formativen Beurteilung

> Mir ist der Entwicklungsprozess wichtiger und darum sind die Aufzeichnungen von uns auch wichtig: Wie ist das Kind in die Schule gekommen und was hat es für einen Entwicklungsprozess gemacht? Und dies kann die Messung von Bildungsstandards nicht zum Ausdruck bringen (V3).

Allerdings sind Bildungsstandards hilfreich als Orientierungspunkte bei der Klärung der individuellen Entwicklung.

5 Diskussion

5.1 Zusammenführung der Ergebnisse aus den quantitativen und qualitativen Daten

Um die Ergebnisse aus einem Mixed-Methods-Design zusammenzuführen, lässt sich auf die vier Strategien von Caracelli und Greene (1993) zurückgreifen: 1. Transformation der Daten in die jeweils andere Form, 2. Entwicklung einer gemeinsamen Typologie, 3. Fallanalyse extremer Fälle, 4. Daten konsolidieren und gemeinsam auswerten. In unserem Fall ist die Strategie der Entwicklung von Typen oder Kategorien sinnvoll, da die Probanden nur teilweise in beiden Datensätzen enthalten sind. Bei unserer Studie handelt es sich um Typen hinsichtlich der Anwendung von Elementen einer formativen Beurteilung. Zeigen sich in den beiden Analysen ähnliche Typen mit ähnlichen Elementen? In der Latent Profile Analyse und in der Latent Class Analyse sind die Verhältnisse der Lehrpersonen in den beiden Gruppen etwa 1:5. Sämtliche Kategorien der LPA sind auch Teil der LCA. In der LCA sind noch zusätzliche Kategorien vorhanden und die beiden Gruppen differenzieren stärker. Dies ist vermutlich auf den dichotomen Charakter der Daten (vorhanden/ nicht vorhanden) zuzuführen. In der LPA scheinen die Lehrpersonen die Elemente formativer Beurteilung gesamthaft mehr einzusetzen. Das liegt sicherlich auch daran, dass in den Interviews nicht explizit nach Aspekten gefragt wurde, sondern analysiert wurde, was die Lehrpersonen unaufgefordert nannten. Das LCA-Modell vertieft nicht nur das LPA-Modell, sondern bestätigt im grossen Ganzen auch die beiden Typen. In einem nächsten Schritt haben wir versucht, die Ergebnisse interpretativ zu generalisieren. Ausgegangen sind wir dazu von den beiden Typen sowie dem in der formativen Beurteilung oft verwendeten Kreislaufmodell (Oggenfuss et al. 1995; Smit 2009; Greenstein 2010) mit den vier Elementen: Ziele (in der Unterrichtplanung festlegen), Beobachten, Evaluieren, Rückmelden.

Für die Darstellung der Synthese wurde eine Tabelle gewählt (Tab. 2), in der jeweils die Merkmale für die beiden Gruppen traditionell und erweitert zusammenfassend dargestellt sind. Die Synthese wurde nach den vier Elementen des Kreislaufmodells gegliedert. In der Diskussion wird die Bedeutung der Ergebnisse für Ausbildung und Forschung erörtert.

Tab. 2 Synthese der beiden LCA

Beurteilung	Traditionell	Erweitert
Quantitative Daten	Portfolio und schriftlicher Lernbericht werden selten oder nie eingesetzt. Lernstand demonstrieren lassen, gezielt beobachten und auch festhalten, Lernentwicklung aufzeigen und individuell beurteilen wird zwischen selten und manchmal benutzt	Portfolio und schriftlicher Lernbericht werden manchmal oder selten eingesetzt. Lernstand demonstrieren lassen und gezielt beobachten wird zwischen manchmal und oft benutzt. Lernentwicklung aufzeigen, individuell beurteilen und Beobachtung schriftlich festhalten wird oft eingesetzt
Qualitative Daten	Portfolio, Lernbericht, Selbstund Peerbeurteilung, individuelle Beurteilung und gezieltes und dokumentiertes Beobachten sind eher unwahrscheinlich in dieser Gruppe anzutreffen. Nur formative Lernkontrollen haben eine kleine Wahrscheinlichkeit bei dieser Gruppe eingesetzt zu werden	Peer-Beurteilung ist eher unwahrscheinlich, Portfolio und Lernbericht schreiben weist eine kleine Wahrscheinlichkeit auf. Selbstbeurteilung, und gezielte Beobachtung weisen eine mittlere Wahrscheinlichkeit auf. Eine höhere Wahrscheinlichkeit zeigt sich bei formativer Lernkontrolle, dem Festhalten von Beobachtungen, individueller Beurteilung und dem Planen auf Grund von Ergebnissen aus einer Beurteilung
Synthese	Die Lektionsplanung basiert eher nicht auf Beobachtungen und Lernkontrollen. Lediglich formative Lernkontrollen kommen in einem begrenzten Umfang vor. Beobachtungen werden nicht systematisch durchgeführt und festgehalten. Entsprechend werden diese auch nicht genutzt für individuelle Rückmeldegespräche in denen die Lernentwicklung aufgezeigt wird. Ein für diesen Zweck geeignetes Portfolio kommt ebenfalls nicht zum Einsatz. Selbstbeurteilung durch die Lernenden wird auch kaum genutzt. Lehrpersonen in dieser Gruppe setzen Elemente einer formativen Beurteilung nur wenig ein	Die Lektionsplanung basiert auf Beobachtungen und Lernkontrollen. Das Vorwissen wird geklärt. Beobachtungen werden gezielt geplant und schriftlich dokumentiert. Rückmeldungen im Sinne eines Aufzeigens des Lernstandes und der individuellen Entwicklung finden statt. Dabei kommen Portfolios kaum zum Einsatz. Selbstbeurteilung der Lernenden wird teilweise genutzt. Lehrpersonen in dieser Gruppe setzen bestimmt Elemente einer formativen Beurteilung regelmässig und andere gelegentlich ein

5.2 Diskussion

Die Forschungsfragen konnten mit Hilfe der Datenanalysen geklärt werden. Frage 1: Das Konstrukt „formative Beurteilung“ liess sich in der Stichprobe mit den theoretisch vermuteten Elementen nachweisen (Konstruktvalidität). Es erklärt einen substantiellen Teil der Unterschiede in den Daten zur Beurteilungspraxis der Lehrpersonen. Frage 2: Das Modell zeigt auch die vermuteten Zusammenhänge zur Unterrichtplanung und -gestaltung auf. Frage 3: Es zeigen sich sowohl in der Analyse der quantitativen wie auch in den qualitativen Daten zwei Gruppen von Nutzern der formativen Beurteilung: Eine Gruppe mit einer traditionellen und eine mit einer erweiterten Form. Die qualitativen Daten validieren im Sinne einer Triangulation das Messkonstrukt. Wie Maier (2011) schreibt, fehlen im deutschen Raum empirische Studien zur Beschreibung der Praxis formativer Beurteilung an Schulen. Nebst seiner auf der Sekundarstufe 1 angesiedelten Arbeit liegt nun auch eine für die Grundschulstufe respektive Primarstufe vor. Zudem wurde erstmals eine Operationalisierung des Messkonstrukts „formative Beurteilung“ vorgenommen. Vergleicht man die Ergebnisse von Maier mit den vorliegenden, so scheinen die Gymnasiallehrpersonen Elemente einer formativen Beurteilung wie „gezielte Beobachtung“ und „Peer-Beurteilung“ etwas häufiger einzusetzen, das Aufzeigen der Kompetenzentwicklung aber gleich wenig. Es wäre zu klären, ob die Gymnasiallehrpersonen die Items gleich verstehen wie die Grundschullehrpersonen oder ob es schulformbedingte Unterschiede sind.

Da unsere Studie auf Selbstauskünften beruht, bestehen allenfalls gewisse Verzerrungen bei den Fragebogendaten. Den Items liegt auch ein gewisser Interpretationsbedarf inne. Auch bei den Interviewdaten kann auf Grund der offenen Fragestellungen nicht davon ausgegangen werden, dass sämtliche Beurteilungspraktiken erwähnt wurden. Insofern könnten auch die rund 20 % der Gruppe mit einer erweiterten Beurteilung eine Unterschätzung der tatsächlichen Anzahl darstellen. Die Triangulation der beiden Datenquellen weist jedoch auf eine valide Erfassung des Konstrukts hin. Zu erwähnen ist, dass beim Konstrukt „formative Beurteilung“ im SEM der wichtige Aspekt der Feedback-Qualität nicht miteinbezogen wurde. Bei weiteren Studien wären entsprechende Items zu formulieren und auch die Nutzer des Feedbacks miteinzubeziehen, bspw. mit einem Schülerfragebogen (siehe Smit 2009; Harks et al. 2013). Zudem wäre eine Aussensicht durch externe Beobachter oder mittels Videostudien hilfreich, um die bspw. durch Erinnerungseffekte verzerrten Angaben zu prüfen.

5.2.1 Konsequenzen für Aus- und Weiterbildung von Lehrpersonen

Aus der vorliegenden Studie wird deutlich, dass die formative Beurteilung für den grossen Teil der Lehrpersonen in dieser Stichprobe ein wenig umgesetzter Unterrichtsbestandteil ist. Damit wird auch ersichtlich, dass die von einigen deutschsprachigen Autorinnen und Autoren (Carle 2002; Klement 2005; Beutel 2010) geforderte Nutzung einer neuen resp. alternativen Leistungsbewertung zum Zweck einer besseren individuellen Lernunterstützung insbesondere in jahrgangsgemischten Klassen zumindest in den untersuchten Schulen anscheinend noch nicht richtig Fuss gefasst

hat. Allerdings ist die formative Beurteilung auch in der praxisnahen Literatur zum jahrgangsgemischten Unterricht bspw. bei Laging (1999) kaum oder bei Christiani (2005) nur auf wenige Seiten beschränkt ein Thema. Es kann somit vermutet werden, dass auch eher wenig Weiterbildung in diesem Bereich angeboten und nachgefragt wird. Ein entsprechender Ausbildungsbedarf zeigt sich auch in Ländern mit besserem Bekanntheitsgrad von formativer Beurteilung, etwa in den USA (Robinson et al. 2014). Zudem nehmen in vielen Ländern die Lehrpersonen eine Spannung zwischen der Nutzung von formativer Beurteilung und dem Druck, den Ansprüchen und Normen von landesweiten Leistungstests zu genügen, wahr (Smit 2008; Black 2015).

Bezüglich der Ausbildung bleibt abzuwarten, ob es mit einem vermehrten Einbezug von Inhalten zur pädagogischer Diagnostik (Grissemann 2000) in den Ausbildungsstätten in die gewünschte Richtung geht. Müssten nicht die Ausbildungsinhalte im Bereich der Beurteilung weniger auf die Statusdiagnostik und mehr auf die didaktische Gestaltung lernförderlicher Arrangements ausgerichtet sein (Perrenoud 1991), damit es auch zu Veränderungen in der didaktischen Nutzung von formativer Beurteilung in der Praxis zukünftiger Lehrpersonen führt? Dazu gehört auch, dass die Gütekriterien anders zu denken sind: Es besteht weniger der Bedarf an hoch reliablen als an validen Lehrerdiagnosen, welche in Lehr-Lern Gesprächen gewonnen werden (Bohl 2004). Für das selbstgesteuerte Lernen der Schülerinnen und Schüler spielen kontinuierliche, zielunterstützende fachliche und motivationale Rückmeldungen der Lehrpersonen eine bedeutsame Rolle (Boekaerts 1999; Clark 2012). Solches, das Lernen steuernde Feedback basiert auf gezielten wie auch spontanen Beobachtungen im Klassenunterricht, in Schülerarbeitsphasen oder auf der Einsicht in schriftliche Arbeiten von Lernenden. Dieses Feedback sollte abgestimmt sein mit den intendierten Lernzielen der Unterrichtsplanung. In Ergänzung dazu erhält die Beurteilung des Lernstandes oder die Diagnose von Fehlern resp. fehlenden Kenntnissen ihre Reliabilität durch das Zusammenführen der oben erwähnten Informationen aus verschiedenen Quellen und durch wiederholtes Beobachten oder explizites Nachfragen. Diese Informationen können auch summativ genutzt werden, beispielsweise wenn es darum geht, sonderpädagogische Leistungen zu beantragen. Die Ergebnisse sind auch für inklusive Klassen von Bedeutung, da auch hier, wie bei den jahrgangsgemischten Klassen, die individuellen Unterschiede deutlich zu Tage treten und ein grössere Heterogenität der Schülerschaft besteht. Watkins (2007) empfiehlt für inklusive Primarschulen vermehrt formative Beurteilungsformen zu nutzen. Die Beurteilung soll sich dabei weniger an Eingangsdiagnosen sondern vermehrt am Lernprozess im Klassenzimmer orientieren. Nebst normativen Tests der Schulpsychologie können auch kompetenzorientierte Leistungskontrollen mit einer kriterialen Bezugsnorm sowie Selbstbeurteilungen (Was könnte den Schüler/die Schülerin unterstützen sowie motivieren, die Lernlücken zu schliessen?) zu einer Optimierung des Unterstützungsangebots für Sonderbedarf beitragen. Bei grosser Heterogenität wird es schwieriger, die soziale Bezugsnorm anzuwenden und die Beurteilung an der Klassenleistung auszurichten (Hargreaves 2001). Die Lehrperson nutzt dazu die Daten aus verschiedenen Beurteilungssituationen und vergleicht diese mit einem jahrgangsbezogenen Standard (Bourke und Mentis 2014). Damit wird deutlich, dass formative und summative Beurteilungen nicht auf bestimmten Instrumenten beruhen, sondern, dass die beiden Beurteilungsformen unterschiedliche Funktionen aufweisen (Gipps 1994). Sie ergänzen

sich in einer dynamischen Weise und beziehen sowohl Lernprozess und -ergebnis mit ein.

5.2.2 Konsequenzen für die Forschung

Offen bleibt die Frage, inwiefern die formative Beurteilung, so wie sie hier operationalisiert wurde, eine Wirkung auf Schülerleistungen oder -selbstregulative Kompetenzen hat. Zumindest aus Lehrersicht ergeben sich im präsentierten Modell Hinweise zur Wirkung auf letztere. Zu prüfen bleibt weiter, ob das Konstrukt und der Zusammenhang mit den Selbstkompetenzen nur in jahrgangsgemischten Klassen oder generell messbar sind. Gerade der jahrgangsgemischte Unterricht erfordert es ja, dass die Kinder vermehrt selbständig arbeiten und sich gegenseitig Hilfe geben, da die Lehrperson ihre Zeit öfters auf mehrere Jahrgangsstufen verteilen muss. Grundsätzlich geht es aber in Zukunft darum, die Effekte von formativer Beurteilung vermehrt als Gesamtkonstrukt und nicht nur einzelne Aspekte etwa Feedback, Peer-Beurteilung oder Evaluation des Unterrichts gesondert zu prüfen. Insbesondere im französischen Sprachraum besteht diese ganzheitliche Sicht von formativer Beurteilung schon länger (Allal und Lopez Mottier 2005). Dabei wird die Schwierigkeit bestehen bleiben, dass die konkrete Umsetzung der formativen Beurteilung bei jeder Lehrperson wieder ein wenig anders aussieht (Black und Wiliam 1998a). Black (2015) empfiehlt Forschenden, sich in der Folge als gemeinsame Referenz an ein Kreislaufmodell einer Beurteilung (siehe 4.3) zu halten, welches als Teil der Theorie des Lernens und der Selbstregulation betrachtet werden kann (Perrenoud 1998).

Literatur

Achermann, E., & Gehrig, H. (2011). *Altersdurchmischtes Lernen AdL*. Bern: Schulverlag plus.

Allal, L., & Mottier, L. L. (2005). Formative assessment: a review of publications in French. In Organisation for Economic Cooperation and Development (Hrsg.), *Formative assessment: improving learning in secondary classrooms* (S. 241–264). Paris: Organisation for Economic Cooperation and Development.

Baas, D., Castelijns, J., Vermeulen, M., Martens, R., & Segers, M. (2015). The relation between assessment for learning and elementary students' cognitive and metacognitive strategy use. *British Journal of Educational Psychology*, *85*(1), 33–46.

Baeriswyl, F., & Bertschy, B. (2010). Schulische Leistungen kohärent beurteilen – auch bei Inklusionen. *Schweizerische Zeitschrift für Heilpädagogik*, *16*(9/10), 12–19.

Bandura, A. (1989). *Self-regulation of motivation and action through internal standards and goal systems*. New Jersey: Lawrence Earlbaum Associates.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238.

Bazeley, P. (2009). Editorial: integrating data analyses in mixed methods research. *Journal of Mixed Methods Research*, *3*(3), 203–207.

Benjamin, A. (2013). *Formative assessment for English language arts: a guide for middle and high school teachers*. New York: Routledge.

Besser, M., Blum, W., & Klimczak, M. (2013). Formative assessment in everyday teaching of mathematical modelling: implementation of written and oral feedback to competency-oriented tasks. In G. A. Stillman, G. Kaiser, W. Blum, & J. P. Brown (Hrsg.), *Teaching mathematical modelling: connecting to research and practice* (S. 469–478). Dordrecht: Springer Netherlands.

Beutel, S.-I. (2010). Mit Kindern im Gespräch: Lernbegleitung und Leistungsbeurteilung im jahrgangsheterogenen Gruppen. In H. Hahn & B. Berthold (Hrsg.), *Altersmischung als Lernressource* (S. 123–135). Baltmannsweiler: Schneider Verlag Hohengehren.

Black, P. (2015). Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 161–177.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*, 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139–148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box. *Phi Delta Kappan*, *86*(1), 8–21.

Boekaerts, M. (1999). Self-regulated learning: where we are today. *International Journal of Educational research and evaluation*, *31*, 445–457.

Bohl, T. (2004). *Prüfen und Bewerten im Offenen Unterricht*. Weinheim: Beltz.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Hrsg.), *Handbook of causal analysis for social research* (S. 301–328). Dodrecht: Springer Netherlands.

Bourke, R., & Mentis, M. (2014). An assessment framework for inclusive education: integrating assessment approaches. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 384–397.

Bruner, J. S. (1960). *The process of education*. Cambridge: Harvard University Press.

Bürgermeister, A. (2013). *Leistungsbeurteilung im Mathematikunterricht: Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit*. Münster: Waxmann.

Burner, T. (2014). The potential formative benefits of portfolio assessment in second and foreign language writing contexts: a review of the literature. *Studies in Educational Evaluation*, *43*(0), 139–149.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational evaluation and policy analysis*, *15*(2), 195–207.

Carle, U. (2002). Neue Formen der Leistungsdokumentation in heterogenen Lerngruppen. http://www.ganztaegig-lernen.de/media/material/leistungsdokumentation.pdf. Zugegriffen: 07. April 2014.

Christiani, R. (Hrsg.). (2005). *Jahrgangsübergreifend unterrichten*. Berlin: Cornelsen.

Clark, I. (2012). Formative assessment: assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249.

Cornish, L. (2006). *Reaching EFA through multi-grade teaching: issues, contexts and practices*. Armidale: Kardoorair Press.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice*, *6*(1), 101–116.
Creswell, J., & Plano, C. V. (2011). *Designing and conducting mixed methods research* (2. Aufl.). Thousand Oaks: SAGE.
Dick, W. (1996). The dick and carey model: will it survive the decade? *Educational Technology Research and Development*, *44*(3), 55–63.
Dochy, F. (2001). A new assessment era: different needs, new challenges. *Research Dialogue in Learning and Instruction*, *2*, 11–20.
Eckerth, M. (2013). *Formen der Diagnose und Förderung: eine mehrperspektivische Analyse zur Praxis pädagogischer Fachkräfte in der Grundschule*. Münster: Waxmann.
Geiser, C. (2010). *Datenanalyse mit Mplus*. Wiesbaden: Springer VS.
Gipps, C. V. (1994). *Beyond testing: towards a theory of educational assesment*. London: Falmer.
Gläser-Zikuda, M., & Lindacher, T. (2007). Portfolioarbeit im Unterricht-praktische Umsetzung und empirische Überprüfung. In M. Gläser-Zikuda & T. Hascher (Hrsg.), *Lernprozesse analysieren, fördern und beurteilen – Lerntagebuch und Portfolio in Bildungsforschung und Bildungspraxis* (S. 189–204). Bad Heilbrunn: Klinkhardt.
Greenstein, L. (2010). *What teachers really need to know about formative assessment*. Alexandria: ASCD. Grissemann, H. (2000). Pädagogische Diagnostik in der Lehrerbildung an Fachhochschulen. Grundlagen und Systematik pädagogischen Verstehens – eine curriculare Skizze. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *18*(2), 223–231.
Grunder, H.-U., & Bohl, T. (2001). *Neue Formen der Leistungsbeurteilung in den Sekundarstufen I und II*.
Baltmannsweiler: Schneider Verlag Hohengehren.
Guskey, T. R. (2007). Closing achievement gaps: revisiting benjamin S. bloom's "learning for mastery". *Journal of advanced academies*, *19*(1), 8–31.
Hargreaves, E. (2001). Assessment for learning in the multigrade classroom. *International Journal of Educational Development*, *21*, 553–560.
Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2013). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology*, *34*(3), 1–22.
Harlen, W. (2007). The impact of summative assessment on children, teaching, and the curriculum. In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann, & R. Schages (Hrsg.), *Qualität von Grundschulunterricht* (S. 51–65). Wiesbaden: Springer VS.
Hattie, J. (2002). Classroom composition and peer effects. *International Journal of Educational Research*, *37*, 449–481.
Hattie, J. (2008). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. Heritage, M. (2007). Formative assessment: what do teachers need to know and do? *Phi Delta Kappan*, *89*(2), 140–145.
Houtveen, A. A. M., Booij, N., Jong, R. de, & Grift, W. J. C. M. van de (1999). Adaptive instruction and pupil achievement. *School Effectiveness and School Improvement*, *10*(2), 172–192.
Ingenkamp, K. (1995). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz. Ingenkamp, K. (2005). *Lehrbuch der Pädagogischen Diagnostik*. Weinheim: Beltz.
Katz, L. G., Evangelou, D., & Allison Hartman, J. (1990). *The case for mixed-age grouping in early education*. Washington, D.C: National Association for the Education of Young Children.
Keeley, P. (2008). *Science formative assessment: 75 practical strategies for linking assessment, instruction, and learning*. Thousand Oaks: Corwin Press.
Klement, K. (2005). Alternative Leistungsbeurteilung in heterogenen Lerngruppen – Methodische Trägerkriterien als Voraussetzung pädagogischer Beurteilungsformen. *Erziehung & Unterricht*, *155*(5–6), 423–435.
Kuhl, P., Felbrich, A., Richter, D., Stanat, P., & Pant, H. A. (2013). Die Jahrgangsmischung auf dem Prüfstand: Effekte jahrgangsübergreifenden Lernens auf Kompetenzen und sozio-emotionales Wohlbefinden von Grundschülerinnen und Grundschülern. In R. Becker & A. Schulze (Hrsg.), *Bildungskontexte* (S. 299–324). Wiesbaden: Springer.
Laging, R. (Hrsg.) (1999). *Altersgemischtes Lernen in der Schule*. Baltmannsweiler: Schneider Verlag Hohengehren.
Lang, E., Grittner, F., Rehle, C., & Hartinger, A. (2010). Das Heterogenitätsverständnis von Lehrkräften im jahrgangsgemischten Unterricht der Grundschule. In J. Hagedorn, V. Schurt, C. Steber, & W. Waburg (Hrsg.), *Ethniztät, Geschlecht, Familie und Schule* (S. 315–331). Wiesbaden: Springer VS.

Little, A. W. (2001). Multigrade teaching: towards an international research and policy agenda. *International Journal of Educational Development*, *21*, 481–497.
Little, A. W. (2007). Multigrade lessons for EFA: a synthesis. In A. W. Little (Hrsg.), *Education for all and multigrade teaching* (S. 301–348). Dordrecht: Springer Netherlands.
Maier, U. (2010). Formative Assessment – Ein erfolgsversprechendes Konzept zur Reform von Unterricht und Leistungsmessung? *Zeitschrift für Erziehungswissenschaft*, *13*(2), 293–308.
Maier, U. (2011). Formative Leistungsdiagnostik in der Sekundarstufe 1 – Befunde einer quantitativen
Lehrerbefragung zu Nutzung und Korrelaten verschiedener Typen formativer Diagnosemethoden in Gymnasien. *Empirische Pädagogik*, *25*(1), 25–46.
Marley, S. C., Szabo, Z., Levin, J. R., & Glenberg, A. M. (2011). Investigation of an activity-based text-processing strategy in mixed-age child dyads. *The Journal of Experimental Education*, *79*(3), 340–360. Mayring, P. (2000). Qualitative Inhaltsanalyse. Forum: Qualitative Sozialforschung, 1(2). http://www.qualitative-research.net/index.php/fqs/article/view/1089/2385. Zugegriffen: 14. Aug 2016.
McMillan, J. H. (2010). The practical implications of educational aims and contexts for formative assessment. In H. Andrade & G. Cizek (Hrsg.), *Handbook of formative assessment* (S. 41–58). New York: Routledge.
Miller, B. A. (1991). A review of the qualitative research on multigrade instruction. *Journal of Research in Rural Education*, *7*(2), 3–12.
Miller, D., & Lavin, F. (2007). 'But now I feel I want to give it a try': formative assessment, self-esteem and a sense of competence. *The Curriculum Journal*, *18*(1), 3–25.
Müller, R., Keller, A., Kerle, U., Raggl, A., & Steiner, E. (Hrsg.) (2011). *Schule im alpinen Raum*. Innsbruck: Studienverlag.
Mulryan-Kyne, C. (2007). The preparation of teachers for multigrade teaching. *Teaching and Teacher Education*, *23*(4), 501–514.
Muthén, L. K., & Muthén, B. O. (2012). Mplus: statistical analysis with latent variables: User's Guide (7. Aufl.). Los Angeles: Muthén & Muthén.
Nicol, D. J., & Macfarlane, D. D. (2006). Formative assessment and self regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.
Nührenbörger, M., & Steinbring, H. (2009). Unterrichtsgespräche zwischen Schülern und Lehrkräften in jahrgangsgemischten Kleingruppen. *Journal of Mathematics Teacher Education*, *12*(2), 111–132.
Oggenfuss, F., Spitzer-Feser, B., Theiler, P., & Vögeli-Mantovani, U. (1995). *Eine Beurteilung, die weiter hilft*. Ebikon: Zentralschweizerischer Beratungsdienst für Schulfragen.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educational Research Review*, *9*(0), 129–144.
Perrenoud, P. (1991). Formative Schülerbeurteilung: Welcher Platz in der Didaktik? *Beiträge zur Lehrerbildung*, *9*(3), 309–329.
Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 85–102.
Peschel, F. (2005). *Offener Unterricht*. Baltmannsweiler: Schneider Verlag Hohengehren. Popham, W. J. (2008). *Transformative assessment*. Alexandria: ASCD.
Raggl, A. (2011). Altersgemischter Unterricht in kleinen Schulen im alpinen Raum. In R. Müller, A. Keller, U. Kerle, A. Raggl, & E. Steiner (Hrsg.), *Schule im alpinen Raum* (S. 231–305). Innsbruck: Studienverlag.
Reusser, K. (1995). Lehr-Lernkultur im Wandel: Zur Neuorientierung in der kognitiven Lernforschung. In
R. Dubs & R. Dörig (Hrsg.), *Dialog Wissenschaft und Praxis* (S. 164–190). St. Gallen: Institut für Wirtschaftspädagogik IWP.
Ricken, G. (2007). Aufgaben der sonderpädagogischen Diagnostik. In K. Salzberg-Ludwig & E. Grüning (Hrsg.), *Pädagogik für Kinder und Jugendliche in schwierigen Lern und Lebenssituationen* (S. 151–164). Stuttgart: Kohlhammer.
Robinson, J., Myran, S., Strauss, R., & Reed, W. (2014). The impact of an alternative professional development model on teacher practices in formative assessment and student learning. *Teacher Development*, *18*(2), 141–162.
Roos, M. (2001). *Ganzheitliches Beurteilen und Fördern in der Primarschule*. Chur: Rüegger.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.
Scheerens, J. (2000). *Improving school effectiveness*. Paris: United Nations Educational, Scientific and Cultural Organization.
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation

models: tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, *8*(2), 23–74.
Schneider, M. C., & Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. In H. L. Andrade & G. Cizek (Hrsg.), *Handbook of formative assessment* (S. 251–276). New York: Routledge.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14. Smit, R. (2008). Formative Beurteilung im kompetenzund standardorientierten Unterricht. *Beiträge zur Lehrerbildung*, *26*(3), 383–392.
Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz*. Baltmannsweiler: Schneider Verlag Hohengehren.
Smit, R., & Humpert, W. (2012). Differentiated instruction in small schools. *Teaching and Teacher Education*, *28*, 1152–1162.
Stefanikis, H. E. (2011). *Differentiated assessment. How to assess the learning potential of every student*. San Francisco: Jossey-Bass.
Stein, R. (2005). Individualisierung durch Kooperation – Aufgaben von Sonderpädagogen in der integrierten Förderung. In M. Götz & K. Müller (Hrsg.), *Grundschule zwischen den Ansprüchen der Individualisierung und Standardisierung* (S. 289–295). Wiebaden: SpringerVS.
Stone, S. J. (1996). *Creating the multiage classroom*. Tucson: Good Year Books.
Trautmann, M., & Wischer, B. (2009). Das Konzept der Inneren Differenzierung – eine vergleichende Analyse der Diskussion der 1970er-Jahre mit dem aktuellen Heterogenitätsdiskurs. In M. A. Meyer,
M. Prenzel, & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (S. 159–172). Wiesbaden: Springer VS.
Ullrich, H. (2015). Die nachmoderne Dorfschule. In M. Kraul (Hrsg.), *Private Schulen* (S. 185–201). Wiesbaden: Springer.
Veenmann, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: a bestevidence synthesis. *Review of Educational Research*, *65*(4), 319–381.
VERBI Software. Consult. Sozialforschung (2012). MAXQDA, Referenzhandbuch. http://www.maxqda.de/download/manuals/MAX11_manual_ger.pdf. Zugegriffen: 19. Februar 2015.
Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Hrsg.), *Applied latent class analysis* (S. 89–106). Cambridge: Cambridge University.
Vygotsky, L. S. (1974). *Denken und Sprechen*. Frankfurt a. M.: Fischer. Wagener, M. (2014). *Gegenseitiges Helfen*. Wiesbaden: Springer.
Watkins, A. (Hrsg.). (2007). *Assessment in inclusive settings: key issues for policy and practice*. Odense: European Agency for Development in Special Needs Education.
Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14. Winter, F. (2004). *Leistungsbewertung*. Baltmannsweiler: Schneider Verlag Hohengehren.
Yin, Y., Shavelson, R.J., Ayala, C.C., Ruiz-Primo, M.A., Brandon, P.R., Furtak, E.M., Tomita, M.K., & Young, D.B. (2008) On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change. *Applied Measurement in Education*, *21*(4), 335–359.