Unterrichtswiss (2020) 48:629–661
https://doi.org/10.1007/s42010-020-00083-7

ALLGEMEINER TEIL

Formatives Assessment im alltäglichen Mathematikunterricht von Primarlehrpersonen: Häufigkeit, Dauer und Qualität

Alois Buholzer · Matthias Baer · Sandra Zulliger · Loredana Torchetti · Merle Ruelmann · Andrea Häfliger · Hanni Lötscher

Eingegangen: 25. Oktober 2019 / Überarbeitet: 10. Juli 2020 / Angenommen: 22. Juli 2020 / Online publiziert: 20. August 2020

Zusammenfassung Formatives Assessment umfasst die Gewinnung von diagnostischen Informationen zum Lernen und seinen Ergebnissen sowie deren Nutzung zur Optimierung von Lehr- und Lernprozessen. Das Ziel des Forschungsprojekts bestand darin, zu untersuchen, mit welcher Häufigkeit, mit welcher Dauer und in welcher Qualität Lehrpersonen formatives Assessment informell im alltäglichen Unterricht durchführen. Um diese Forschungsfragen zu klären, wurde in der vierten Primarschulstufe von 52 Lehrpersonen und ihren Schulklassen im Fach Mathematik je eine Doppelstunde zur Einführung in das halbschriftliche Dividieren videografiert. Im Fokus standen fünf Strategien des formativen Assessments: (1) Lernziel klären, (2) Lernstand erfassen (Eliciting-Evidence), (3) Self-Assessment, (4) Peer-Assessment und (5) Feedback-Interaktion. Hinsichtlich der Häufigkeit ergaben die Videoanalysen, dass der überwiegende Teil der teilnehmenden Lehrpersonen alle fünf Strategien nutzt, die Einsatzdauer mit Ausnahme der Feedback-Interaktionen in der Regel jedoch kurz ist. Die Ergebnisse der Qualitätseinschätzung deuten insgesamt auf eine höchstens mittlere, meist aber geringe Qualität der Umsetzung der Strategien hin. Es bestehen Zusammenhänge zwischen der Anwendungsdauer und der Qualitätsausprägung der Strategien. Je länger die Strategien eingesetzt wurden, desto besser wurde ihre Qualität im Rating eingeschätzt. Diese Ergebnisse erweitern den bisherigen Forschungsstand insofern, als erstmals ökologisch valide Aussagen für den deutschen Sprachraum zur Häufigkeit, zur Dauer und zur Qualität von Stra-

A. Buholzer (✉) · S. Zulliger · M. Ruelmann · A. Häfliger · H. Lötscher
Pädagogische Hochschule Luzern, Luzern, Schweiz
E-Mail: alois.buholzer@phlu.ch

M. Baer
Pädagogische Hochschule Zürich, Zürich, Schweiz

Universität Zürich, Zürich, Schweiz

L. Torchetti
Pädagogische Hochschule Bern, Bern, Schweiz

tegien des formativen Assessments im alltäglichen Mathematikunterricht vorgelegt werden. Aus den Befunden werden Schlussfolgerungen für die Aus- und Weiterbildung von Lehrpersonen gezogen.

Schlüsselwörter Formatives Assessment · Lernziele · Lernstanderhebung · Peer-Assessment · Self-Assessment · Feedback-Interaktion · Videostudie · Diagnostik · Formative Beurteilung

1 Einleitung

Mit formativem Assessment werden im Unterricht diagnostische Informationen zum Lernen und seinen Ergebnissen erfasst, die zur Optimierung von Lehr- und Lernprozessen genutzt werden können (Schütze et al. 2018). In Ergänzung dazu fasst summatives Assessment diagnostische Informationen meist am Ende einer Unterrichtseinheit zusammen, damit eine abschließende Einschätzung zur Erreichung der Lernziele vorgenommen werden kann (Black und Wiliam 2009; Pellegrino et al. 2001; Sadler 1989). Hinsichtlich des Planungs- und Formalisierungsgrads kann das formative Assessment variieren. Die Umsetzung kann weitgehend situativ und in Interaktion mit den Schülerinnen und Schülern im alltäglichen Unterricht erfolgen oder im Voraus von der Lehrperson festgelegt und für die ganze Klasse geplant werden (Ruiz-Primo und Furtak 2006). Im vorliegenden Beitrag wird die informelle Umsetzung im alltäglichen Unterricht fokussiert. Die Wirkung eines solchen (informellen) formativen Assessments hängt aktuellen konzeptionellen Überlegungen (Black und Wiliam 2009) und Forschungsbefunden (z. B. Ruiz-Primo und Furtak 2006) zufolge entscheidend mit der Qualität der kognitiven Aktivierung, dem Einbezug der Schülerinnen und Schüler sowie der ineinandergreifenden Verbindung von diagnostischen Informationen und deren adaptiver Nutzung zusammen. Wenn formatives Assessment mit hoher Qualität umgesetzt wird, gilt es als eines der wirksamsten Rahmenkonzepte, um schulisches Lernen zu fördern und positive Effekte sowohl auf kognitive als auch auf metakognitive und motivationale Merkmale der Schülerinnen und Schüler zu erreichen (Decristan et al. 2015).

Trotz des großen Potenzials, das dem informellen formativen Assessment zugeschrieben wird (Ruiz-Primo 2011), liegen im deutschsprachigen Raum zu dessen Häufigkeit und Qualität im alltäglichen Unterricht nur Forschungsergebnisse vor, die auf Selbstauskünften von Lehrpersonen beruhen. Die verfügbaren Befunde zur Häufigkeit sind zudem nicht konsistent. Sie weisen einerseits auf einen häufigen Einsatz hin (Schmidt 2020; Schmidt und Liebers 2015), lassen andererseits aber auch darauf schließen, dass formatives Assessment nur selten einen Bestandteil des Unterrichts bildet (Smit und Engeli 2017). Bezüglich der Qualität des formativen Assessments liegen einzelne Video- und Beobachtungsstudien vor, die insgesamt auf eine eher geringe qualitative Umsetzung hindeuten (Gotwals et al. 2015; Oswalt 2013; Ruiz-Primo und Furtak 2006).

Obwohl somit erste Erkenntnisse vorliegen, mangelt es insbesondere für den deutschsprachigen Raum nach wie vor an Forschungsarbeiten, die das informelle formative Assessment im alltäglichen Unterricht aus einer Außensicht erfassen (Smit

und Engeli 2017). Einen Beitrag zur Schließung dieser Lücke will die vorliegende Teilstudie des Forschungsprojekts TUFA[1] leisten, in der die Praxis des informellen formativen Assessments im alltäglichen Mathematikunterricht durch geschulte unabhängige Beobachterinnen und Beobachter hinsichtlich Quantität und Qualität erfasst und beurteilt wurde.

2 Theoretischer Hintergrund und Forschungsstand zum formativen Assessment

2.1 Formatives Assessment: Begriffsbestimmung

Im deutschen Sprachraum wird formatives Assessment auch „formative Leistungsdiagnostik" (Klauer 2014) oder „formative Beurteilung" (Smit 2008) genannt. Es lenkt den Fokus darauf, „*what* the learner knows, understands or can do" (Hattie 2003, S. 2) und erfüllt die Funktion eines „assessment for learning". Mit formativem Assessment holt die Lehrperson fortlaufend diagnostische Informationen zum Lernprozess ein (formative Diagnostik), um die Schülerinnen und Schüler durch Rückmeldungen (formatives Feedback) bei der Steuerung ihres Lernens zu unterstützen und sie mithilfe einer adaptiven Gestaltung des Unterrichtsangebots individuell begleiten zu können (Bürgermeister und Saalbach 2018; Schmidt und Liebers 2015; Schütze et al. 2018). Formative Diagnostik und formatives Feedback stehen deshalb stets in einem engen Zusammenhang (Bürgermeister und Saalbach 2018). Im Gegensatz dazu erfolgt das summative Assessment, das auch als „assessment of learning" bezeichnet wird, meist als zusammenfassende Beurteilung am Ende einer Unterrichtseinheit, um Informationen zur Zielerreichung einzuholen und/oder Selektionsentscheide zu treffen (Black und Wiliam 2009; Pellegrino et al. 2001; Sadler 1989). Der Hauptunterschied zwischen formativem und summativem Assessment besteht somit in der Nutzung der diagnostischen Informationen und in der Verortung im Lernprozess (Black und Wiliam 2018).

Die dem vorliegenden Beitrag zugrunde gelegte, explizit auf das Lernen ausgerichtete Auffassung von formativem Assessment hebt sich deutlich von einem Verständnis ab, das sich reduktionistisch auf die Auswahl und die Anwendung von diagnostischen Instrumenten bezieht (zur Kritik am eng gefassten Begriff Heritage 2010). Mit dieser erweiterten Konzeption des formativen Assessments einher geht der Grundsatz, „dass Lehrkräfte und Schülerinnen und Schüler als Agierende notwendig sind" (Schütze et al. 2018, S. 701). Formatives Assessment wird somit *nicht für*, sondern *mit* den Schülerinnen und Schülern vorgenommen: „Expert FA [Formative Assessment] is a dialogic process that allows students' ideas to guide instruction and learning" (Gotwals et al. 2015, S. 407). Diese partizipative Ausrichtung fasst Schülerinnen und Schüler als „eigenständige Lernende" (Reusser 1995) auf und bezieht sie aktiv in die Beurteilung von eigenen Lernprozessen oder diejenigen von Mitschülerinnen und Mitschülern (Self- und Peer-Assessments) ein.

[1] „*T*eachers' *U*se of *F*ormative *A*ssessment" – Formativ-diagnostisches Handeln im Unterricht: Voraussetzungen von Lehrpersonen und Effekte bei Schüler/-innen (SNF-Projekt Nr. 100019 169771).

Bell und Cowie (2001) unterscheiden hinsichtlich der Umsetzung zwischen dem formalen oder geplanten formativen Assessment zur Erfassung von diagnostischen Informationen über eine ganze Klasse und dem informellen oder interaktiven formativen Assessment, das diagnostische Informationen situativ in Interaktionen zwischen der Lehrperson und den Schülerinnen und Schülern erfasst. Letzteres wird auch als „on-the-fly formative assessment" bezeichnet (Ruiz-Primo und Furtak 2006; Shavelson et al. 2008). Ruiz-Primo und Furtak (2006) charakterisieren informelles formatives Assessments entsprechend wie folgt: „The framework we propose is based on the idea that informal formative assessment can take place at any level of student-teacher interaction in the course of daily classroom talk, whether whole class, small group, or one-on-one ..., and can help teachers continuously acquire information about their students' level of understanding" (S. 207). Dieses informelle, im Unterrichtsverlauf stattfindende formative Assessment steht im Zentrum des vorliegenden Beitrags.

2.2 Strategien des formativen Assessments im Unterricht

Gemäß der Konzeption von Black und Wiliam (2009) manifestiert sich formatives Assessment im Unterricht, wenn die Lehrperson (1) Lernziel und Beurteilungskriterien mit den Schülerinnen und Schülern teilt, (2) anregende Fragen zum Lernstand stellt (Eliciting-Evidence), die Schülerinnen und Schüler zu (3) Self-Assessments und (4) Peer-Assessments anleitet oder (5) (Fremd-)Beurteilungen mit Lernunterstützung (Feedback-Interaktion) verbindet. Diese fünf „key strategies" ermöglichen gemäss Black und Wiliam (2009) die Klärung dreier Grundfragen zum formativen Assessment: „Wohin geht das Lernen?", „Wo steht die Schülerin oder der Schüler zurzeit?" und „Wie gelangt die Schülerin oder der Schüler zum Ziel?". Hinsichtlich der Wirkungen dieser Strategien sind sowohl die Quantität als auch die Qualität der Umsetzung von Bedeutung: Sie sollen gemäß Wylie und Lion (2013) im Unterricht regelmäßig und in bester Ausprägung eingesetzt werden. Als Kernmerkmale der Qualität gelten diesbezüglich die kognitive Aktivierung der Schülerinnen und Schüler durch vielfältige Anregungen der Lehrperson, ihr aktiver Einbezug in den Assessment-Prozess sowie die ineinandergreifende Verbindung von diagnostischen Informationen und adaptivem Lernangebot (Black und Wiliam 2009; Heritage 2007; Ruiz-Primo und Furtak 2006). Die Qualität ist dann hoch, wenn das formative Assessment bei den Schülerinnen und Schülern Fragen und Erklärungen herausfordert, eigene Beurteilungsprozesse auslöst, Metakognition bewirkt und generell komplexe(re) fachliche, aber auch überfachliche kognitive Aktivitäten anregt. Dazu gehört unter anderem, dass Lehrpersonen die Schülerinnen und Schüler dazu anleiten, ihr Verständnis und ihr Können mit den zu erreichenden Lernzielen bzw. den vorgegebenen Beurteilungskriterien zu vergleichen und daraus abzuleiten, was sie bereits verstanden bzw. nicht verstanden haben und wie sie in ihrem Lernprozess weiterfahren sollen (Gotwals und Birmingham 2016; Ruiz-Primo 2011).

Nachfolgend wird auf die fünf Strategien eingegangen, die den Kodier- und Ratinginstrumenten in der vorliegenden Studie zugrunde liegen.

I Transparente Lernziele und Beurteilungskriterien Diese Strategie zielt darauf ab, zusammen mit den Schülerinnen und Schülern eine Vorstellung darüber zu entwickeln, in welche Richtung sich das Lernen bewegt (Lernziele) und wie überprüft werden kann, ob das, was gelernt werden soll, auch tatsächlich gelernt wurde (Beurteilungskriterien). Eine hohe qualitative Ausprägung dieser Strategie liegt vor, wenn Lernziele und ihre Beurteilungskriterien kognitiv aktivierend und anhand von konkreten Beispielen (Produkte, Ergebnisse) besprochen werden, Zusammenhänge mit dem Vorwissen und dem vorangegangenen Unterricht transparent gemacht werden (Kobarg und Seidel 2003) und die Lernziele und Beurteilungskriterien in eine Alltagssituation oder eine Problemstellung eingebettet sind (Wylie und Lyon 2013). Kennzeichnend für eine hohe Qualität ist außerdem, dass Lernziele und Beurteilungskriterien präzise formuliert und adressatengerecht kommuniziert (Sadler 1989) und als Vorschau oder Rückblick auf Lernprozesse genutzt werden (Trepke et al. 2003).

II Eliciting-Evidence (Ermittlung des Lernstands) Mit dieser Strategie verfolgt die Lehrperson das Ziel, Einsicht in die Konzepte und die Vorstellungen der Schülerinnen und Schüler und damit in ihren aktuellen Lernstand zu gewinnen (Ruiz-Primo 2011; Schmidt 2020). Im informellen formativen Assessment findet eine solche Ermittlung des Lernstands häufig interaktiv im Unterrichtsgespräch statt. Eine hohe Qualität ist bei der Umsetzung dieser Strategie dann erkennbar, wenn die Schülerinnen und Schüler durch Fragen und Impulse dazu herausgefordert werden, eigene Ideen, Konzepte und Lösungen zu entwickeln. Zu diesem Zweck soll ihnen ausreichend Zeit zum Nachdenken eingeräumt werden, damit sie ihre Denkwege und Überlegungen ausführlich darlegen und erklären können (Harlen 2007; Heritage 2010; Lotz 2016). Ebenfalls unabdingbar sind zudem eine aufmerksame Überwachung der Lernprozesse und die Überprüfung der Lernergebnisse (Produkte) im Hinblick auf die angestrebten Ziele (Cizek 2010; Harlen 2007).

III Self-Assessment Mithilfe dieser Strategie sollen die Schülerinnen und Schüler dabei unterstützt werden, ihr Verständnis und ihr Können selbst zu beschreiben, ihren Lernstand mit den zu erreichenden Lernzielen bzw. den von der Lehrperson vorgelegten Beurteilungskriterien zu vergleichen und daraus abzuleiten, was sie verstanden bzw. nicht verstanden haben und wie sie im Lernen/Problemlösen weiterfahren wollen. Eine hohe Qualität manifestiert sich bei dieser Strategie, wenn das Self-Assessment anregend gestaltet ist und zu vertiefter Selbstreflexion anregt (Wylie und Lyon 2013). Außerdem ist es wichtig, dass die Erkenntnisse aus dem Self-Assessment systematisch für den Unterricht und die nächsten Lernschritte genutzt werden (Andrade 2010) und dass die Lehrperson das Self-Assessment klar und mit Bezug auf die Lernziele anleitet und die Durchführung sorgfältig begleitet (Harris und Brown 2013).

IV Peer-Assessment Mit dieser Strategie wird intendiert, dass sich die Schülerinnen und Schüler mit dem Lernen und den Lernergebnissen ihrer Peers befassen, indem sie sich auf der Grundlage des Lernziels und den vorgegebenen Beurteilungskriterien gegenseitig Rückmeldungen zu ihrem Verständnis und ihrem Können

geben und gemeinsam über weitere Lösungswege nachdenken (Topping 2010). Eine qualitätsvolle Umsetzung zeigt sich ähnlich wie beim Self-Assessment daran, dass der Austausch unter den Peers zu einer vertieften Reflexion von eigenen und fremden Lernprozessen anregt und die gewonnenen Einsichten zur Optimierung von Lehr- und Lernprozessen genutzt werden. Wichtig im Hinblick auf die Sicherstellung einer hohen Qualität ist außerdem, dass die Lehrperson das Peer-Assessment sorgfältig anleitet und begleitet.

V Feedback-Interaktion Das Ziel der häufig in mündlicher Interaktion (Ruiz-Primo und Furtak 2006) umgesetzten Strategie besteht darin, die Differenz zwischen Lernstand und Lernziel mithilfe von Feedbacks der Lehrperson sowie der Mitschülerinnen und Mitschüler zu überwinden. Hohe Qualität lässt sich hierbei konstatieren, wenn der Sinn oder die Ziele von Aufgaben und Tätigkeiten verdeutlicht werden (Feed Up), der Lernstand beschrieben und beurteilt wird (Feed Back) sowie ausgehend vom aktuellen Stand die nächsten Lernschritte und Herausforderungen festgelegt werden (Feed Forward) (Hattie und Timperley 2007). Außerdem sollen Feedbacks spezifisch und aktivierend für das Lernen der Schülerinnen und Schüler formuliert sein. Von Relevanz sind zudem die Anpassung der Feedbacks an den Kenntnisstand der Schülerinnen und Schüler, die Berücksichtigung von motivationalen und emotionalen Aspekten (Hattie und Timperley 2007; Kluger und DeNisi 1996) sowie die Wechselseitigkeit in den Feedback-Interaktionen und somit die dialogische Gestaltung des Assessments (Ruiz-Primo und Furtak 2006).

Nach Hattie und Timperley (2007) lässt sich die mündliche Ermittlung des Lernstands ebenfalls als Feedback auffassen. Auch Heritage (2010) zufolge gehören die Ermittlung des Lernstands und das darauffolgende Feedback zusammen, weshalb die Strategien II und V im Gegensatz zur Konzeption von Black und Wiliam (2009) nicht als zwei separate Strategien des formativen Assessments aufgefasst werden.

2.3 Empirische Befunde zur Quantität und Qualität der Umsetzung formativen Assessments

Empirische Befunde zur Quantität des formativen Assessments beziehen sich darauf, wie oft und mit welcher Dauer Lehrpersonen es im Unterricht einsetzen. Gemäß den Ergebnissen von Bürgermeister (2014) kommt verbale Leistungsbeurteilung in Form von Feedbacks in der Sekundarstufe I häufig vor, während Peer- und Self-Assessments mit einer Häufigkeit zwischen „nie“ und „manchmal“ durchgeführt werden. Lehrpersonen der Grundschulstufe stimmten in der Studie von Schmidt (2020) überwiegend der Aussage zu, dass sie in jeder Stunde Lernziele nennen und besprechen. Die meisten Befragten berichteten zudem, mit mündlichen Feedbacks das Lernen der Schülerinnen und Schüler zu unterstützen. Etwas weniger Zustimmung erfuhren die Items zum Einsatz von Self- und Peer-Assessments. In der Untersuchung von Maier (2011) gab die Mehrheit der teilnehmenden Gymnasiallehrpersonen im Gegensatz dazu an, Peer-Assessments im Unterricht mit der Frequenz „öfter“ bis „regelmäßig“ einzusetzen (58 %).

Die in diesen Studien insgesamt vorgenommene weitgehend positive Einschätzung hinsichtlich der Häufigkeit des Einsatzes formativer Assessments teilen Smit

und Engeli (2017) in Anbetracht der Auswertung ihrer Befragung nicht. Sie unterscheiden aufgrund einer Latent Profile Analysis vielmehr zwischen zwei Gruppen von Nutzerinnen und Nutzern, wobei Lehrpersonen der Gruppe „Erweiterte Beurteilung" ($N = 14$) im Gegensatz zu Lehrpersonen der Gruppe „Traditionelle Beurteilung" ($N = 61$) deutlich untervertreten sind. Ihre Ergebnisse resümierend halten Smit und Engeli (2017) fest, „dass die formative Beurteilung ein wenig umgesetzter Unterrichtsbestandteil" (S. 298) ist. Auch Cheng und Wang (2007) gelangen zum Schluss, dass die „traditionelle" Leistungsbeurteilung nach wie vor dominant ist und die Schülerinnen und Schüler nur selten in Prozesse des formativen Assessments einbezogen werden.

Die Frage, mit welcher *zeitlichen Dauer* Strategien des informellen formativen Assessments im Unterricht eingesetzt werden, ist zurzeit noch unbeantwortet. Die vorliegenden Selbsteinschätzungen von Lehrpersonen erlauben keine genaue Bestimmung der Dauer von selbst berichteten Aktivitäten und auch in Beobachtungsstudien zum informellen formativen Assessment finden sich bislang keine Angaben zur Einsatzdauer (Oswalt 2013). Auch diese Studien beschränken sich darauf, die Anzahl der registrierten Strategien auszuzählen und auszuwerten (Gotwals et al. 2015; Ruiz-Primo und Furtak 2006). Eine Videostudie von Krammer (2009) zur individuellen Lernunterstützung kommt zum Schluss, dass pro Mathematiklektion 10,47 min (oder 23 % der Unterrichtszeit) für die inhaltliche, mathematikbezogene Unterstützung (vergleichbar mit Feedback-Interaktion) aufgewendet wurden.

Was die *Qualität von formativen Assessments* betrifft, gelangen empirische Studien, die auf selbst berichtete Daten zurückgriffen, zu positiven Ergebnissen (Altmann et al. 2010; Bürgermeister 2014; Maier 2011; Schmidt 2020). In Beobachtungs- und Videostudien, die auf einer unabhängigen Außenperspektive beruhen, wird die Qualität hingegen weniger hoch beurteilt. Gotwals et al. (2015) beispielsweise fassen die Ergebnisse ihrer Untersuchung wie folgt zusammen: „The teachers in our study, regardless of discipline, did not always (or often) demonstrate high-quality FA practices; however, many of them showed proficiency in specific aspects, such as questioning types, elicitation strategies, and feedback" (S. 419). Bei Oswalt (2013) liegen die Qualitätseinschätzungen zu den Strategien *Eliciting Evidence* und *Feedback* knapp im mittleren Qualitätsbereich, die Strategien *Lernziele*, *Self-Assessment* und *Peer-Assessment* jedoch eher im unteren Qualitätsbereich.

Weitere Befunde zur Qualität finden sich in Forschungsarbeiten zur Wirksamkeit einzelner Strategien auf die Leistungsentwicklung. So sind nach Hattie (2016) Lernziele insbesondere dann leistungssteigernd, wenn sie mit den Schülerinnen und Schülern geteilt werden sowie adaptiv formuliert und an das Vorwissen und die Vorerfahrungen angepasst sind. Gemäß den Ergebnissen von Ruiz-Primo und Furtak (2006) geht eine zyklisch aufgebaute Ermittlung des Lernstands mit besseren Leistungen in einem nachfolgenden Test einher. In dieser Studie riefen die teilnehmenden Lehrpersonen in einem *ESRU*-Zyklus jeweils eine Antwort einer Schülerin oder eines Schülers hervor (*E*liciting), indem sie beispielsweise eine Begründung oder eine Analyse verlangten. Die Antwort (*S*tudent *R*esponse) wurde danach aufgegriffen, paraphrasiert und für den weiteren Unterrichtsverlauf (*U*sing) genutzt. Brookhart et al. (2010) wiederum gelangen zum Ergebnis, dass es sich positiv auf die Leistung auswirkt, wenn die Schülerinnen und Schüler umfangreiches, spezi-

fisches und sofortiges Feedback erhalten. Hinsichtlich der Wirksamkeit von Peer-Feedbacks als Element von Peer-Assessments ist nach Strijbos et al. (2010) entscheidend, ob der Peer als Expertin bzw. Experte wahrgenommen wird und ob das Peer-Feedback konkrete Hinweise zur Verbesserung von Fehlern enthält. In Bezug auf Self-Assessments lassen sich gemäß den Befunden von Panadero et al. (2016) Wirkungen auf die schulische Leistung und insbesondere auf selbstregulatorische und metakognitive Kompetenzen nachweisen.

Insgesamt fallen die berichteten Studienergebnisse zur Frage, wie häufig und in welcher Qualität Strategien des formativen Assessments im Unterricht eingesetzt werden, uneinheitlich aus. Dafür sind vor allem methodische Gründe wie die unterschiedlichen Operationalisierungen des Konstrukts Formatives Assessment oder das Erhebungsformat verantwortlich. Insbesondere die Forschungsarbeiten zum (informellen) formativen Assessment im deutschsprachigen Raum basieren auf Selbstauskünften, was infolge von Milde-Effekten möglicherweise zu Verzerrungen der Urteile geführt haben könnte (Döring und Bortz 2016; Smit und Engeli 2017). Die wenigen verfügbaren Videostudien wiederum liefern zwar eine unabhängige Außenperspektive, stammen jedoch allesamt aus dem angloamerikanischen Kontext, was die Übertragbarkeit der Ergebnisse auf die Schulpraxis im deutschsprachigen Raum infrage stellt (Stigler und Hiebert 1999). Zudem beziehen sie sich auf verschiedene Fächer und Unterrichtsinhalte, wodurch die Vergleichbarkeit der Ergebnisse zusätzlich erschwert wird. Diese Einschränkung ist umso gravierender, als Hinweise zur Domänenspezifität des formativen Assessments vorliegen (Bürgermeister 2014; Maier 2011). Hinzu kommt, dass in diesen Videostudien die Dauer des Einsatzes der Strategien nicht erfasst wurde. Deshalb sind zurzeit auch keine empirisch gesicherten Informationen darüber verfügbar, inwiefern die Dauer von formativem Assessment mit der Qualität zusammenhängt.

3 Fragestellungen

Wie der Überblick über den aktuellen Stand der Forschung zeigt, sind bislang sehr wenige Studien verfügbar, in deren Rahmen unabhängige Beobachterinnen und Beobachter die Praxis des informellen formativen Assessments im alltäglichen Unterricht hinsichtlich ihrer Häufigkeit, Dauer und Qualität erfasst und eingeschätzt haben. Solche Studien wären jedoch insbesondere für den deutschsprachigen Raum unabdingbar. Denn aus validen kontextspezifischen Erkenntnissen ließen sich nicht nur weiterführende Rückschlüsse für künftige Forschungsprojekte, sondern auch praxisbezogene Anknüpfungspunkte für die Aus- und Weiterbildung von Lehrpersonen ziehen (Widmer-Wolf et al. 2014).

Vor dem Hintergrund dieses Forschungsdesiderats wird nachfolgend den folgenden drei Forschungsfragen nachgegangen:

1. Wie hoch ist der *Anteil* der untersuchten Lehrpersonen, die im alltäglichen Unterricht Strategien des informellen formativen Assessments nutzen, und mit welcher *zeitlichen Dauer* setzen sie die fünf Strategien des formativen Assessments ein?

2. Mit welchen *Qualitätsausprägungen* wenden die Lehrpersonen die Strategien des informellen formativen Assessments im alltäglichen Unterricht an?
3. Inwieweit bestehen *Zusammenhänge zwischen der Dauer und der Qualität* des Einsatzes der fünf Strategien des formativen Assessments?

4 Methode

Zur Klärung der Fragestellungen zur Häufigkeit, Dauer und Qualität von formativem Assessment wurden in der vorliegenden Studie Videodaten herangezogen. Anders als selbst berichtete Einschätzungen ermöglichen diese eine Außenperspektive auf den Unterricht, die zwar ebenfalls nicht gänzlich objektiv, jedoch frei von Wahrnehmungen und Interpretationen der beteiligten Schülerinnen und Schüler wie auch der Lehrperson ist (Clausen 2002) und auf standardisierten Instrumenten beruht (Hugener 2008).

4.1 Stichprobe

Die Stichprobe der Gesamtstudie umfasste 52 Lehrpersonen aus der Zentralschweiz, die zum Zeitpunkt der Datenerhebung Klassen der vierten Primarschulstufe in Mathematik und anderen Fächern unterrichteten: 40 Lehrerinnen und 12 Lehrer im durchschnittlichen Alter von 36 Jahren ($SD = 10{,}6$; Range: 25–60 Jahre) mit einer mittleren Berufserfahrung von 10,6 Jahren ($SD = 10{,}3$; Range: 1,5–39,0 Jahre), die in ihren Klassen wöchentlich durchschnittlich 22,1 Unterrichtsstunden von je 45 min Dauer unterrichteten ($SD = 5{,}1$; Range: 4–29 Unterrichtsstunden). Die Lehrpersonen wurden angefragt, eine Einführungsdoppelstunde zum halbschriftlichen Dividieren durchzuführen. Bei der Anfrage wurde ihnen ein mündliches oder schriftliches Feedback zur videografierten Einführungsdoppelstunde angeboten. Da die Lehrpersonen einer Ausschreibung der Pädagogischen Hochschule gefolgt waren, ist von einer positiven Selektion auszugehen. Die Stichprobe der 634 teilnehmenden Schülerinnen und Schüler setzte sich aus je 315 Mädchen und Jungen zusammen; bei vier weiteren lagen keine Angaben zum Geschlecht vor. 38 % der Schülerinnen und Schüler sprachen mit mindestens einem Elternteil eine andere Sprache als (Schweizer-)Deutsch. Das Durchschnittsalter betrug 10,5 Jahre ($SD = 0{,}49$; Range: 9–12,6 Jahre). Sowohl von den Lehrpersonen als auch von den Erziehungsberechtigten der involvierten Schülerinnen und Schüler wurde für die Teilnahme eine aktive Zustimmung eingeholt. 54 von insgesamt 711 Erziehungsberechtigten lehnten eine Teilnahme ihres Kindes ab. Für die Videoaufnahmen wurden die betreffenden Schülerinnen und Schüler deshalb so in der Klasse platziert, dass sie von der Kamera nicht aufgenommen wurden.

4.2 Durchführung der Videostudie

Von jeder der 52 teilnehmenden Lehrpersonen wurde eine Doppelstunde Mathematikunterricht videografiert. Um vergleichbare Rahmenbedingungen sicherzustellen, wurde das Unterrichtsthema „Einführung ins halbschriftliche Dividieren“ vorgege-

ben. Dies erfolgte unter der Annahme, dass der gefilmte Unterricht repräsentativ für den Mathematikunterricht der Lehrpersonen sein dürfte, da subjektive Theorien und Skripts zum Unterrichten als relativ stabil angesehen werden können (Groeben et al. 1988; Pianta und Hamre 2009). Bei der halbschriftlichen Division handelt es sich um einen Lerninhalt aus dem Mathematikcurriculum der vierten Primarklasse in der (Deutsch-)Schweiz.

Die Videoaufnahmen erfolgten wie in den Videostudien von Klieme et al. (2006) und der VideA-Studie von Krammer und Hugener (2014) standardisiert nach einem Kameraskript (Petko 2006), in dem die organisatorischen und technischen Vorbereitungen für die Aufnahmen, die Kameraposition und die Kameraführung detailliert festgehalten waren. Die Videoaufzeichnungen wurden von Mitgliedern der Forschungsgruppe und geschulten studentischen Hilfskräften durchgeführt, wobei die Kamera stets der Lehrperson folgte. Die Lehrperson erhielt zusätzlich ein Ansteckmikrofon, damit ihre sprachlichen Aussagen, auch in Situationen, in denen sie sich mit den Schülerinnen und Schülern im Flüsterton unterhielt, aufgezeichnet werden konnten.

4.3 Entwicklung und Aufbau der Kodier- und Ratinginstrumente für die Videoanalyse

Beobachtungsverfahren, die auf Kodierungen beruhen, eignen sich für die Erfassung des Auftretens, der Häufigkeit und der Dauer eines Merkmals, während sich mithilfe von Schätzverfahren, das heißt Ratings, die Qualität eines Untersuchungsgegenstands erfassen lässt, indem beobachtbare Indikatoren festgelegt und deren Ausprägungen zu einem Gesamteindruck zusammengefasst werden (Pauli 2012). Für die Ermittlung der Häufigkeit und der Dauer (Kodierung) und die Qualitätsbeurteilung (Rating) der in Abschn. 2 erläuterten fünf Strategien des formativen Assessments in den Videoaufnahmen wurden zwei Instrumente entwickelt, pilotiert und eingesetzt (Tab. 1 und 2). Die Notwendigkeit einer Neuentwicklung bestand deshalb, weil die zur Beobachtung von formativem Assessment im Unterricht bereits verfügbaren Instrumente (Gotwals et al. 2015; Oswalt 2013) nicht dafür ausgelegt sind, die Häufigkeit *und* die Qualität zu erfassen. Zur Ausarbeitung der Kodier- und Ratinginstrumente wurden Publikationen der videogestützten Unterrichtsforschung (u. a. TIMS-Studie: Klieme et al. 2006; PERLE-Studie: Lotz 2016; Lotz et al. 2011; IPN-Videostudie Physik: Seidel et al. 2003) herangezogen. Ihre Entwicklung erfolgte theorie- und datengeleitet (Pauli 2012). Die Bestimmung der Kategorien der beiden Instrumente wurde mit dem aus der Inhaltsanalyse stammenden induktiv-deduktiven Vorgehen vorgenommen (Hugener 2006), indem zuerst auf der Basis der Literatur die Oberkategorien (Strategien zum formativen Assessment) und die untergeordneten Facetten/Kategorien bzw. Items deduktiv gesetzt wurden. Die Ausarbeitung, die Beschreibung und die Abgrenzung der Kategorien erfolgten danach induktiv am Videomaterial. Probekodierungen und Proberatings, Besprechungen von dabei auftretenden Schwierigkeiten und die differenzierenden und präzisierenden Anpassungen der Kategoriensysteme wurden rekursiv durchgeführt.

Sowohl für die Häufigkeit als auch für die Qualität von formativem Assessment wurde wie in anderen Studien (TIMSS, PERLE) ein mehrschichtiges, hierarchisch

Tab. 1 Kodierinstrument zur Erfassung der fünf Strategien von formativem Assessment in den videografierten Doppelstunden zur Einführung des halbschriftlichen Dividierens

Strategie	Facette	Kategorien	Cohens Kappa
I Lernziel (LZ) Nennung und Besprechung des Lernziels	A Lernziel	Die Lehrperson ... A1 begründet, warum es wichtig ist, das Lernziel zu bearbeiten. A2 ordnet das Lernziel in die vorangehenden Unterrichtseinheiten oder Wissensbestände ein. A3 bespricht das Lernziel der Lektion oder der Unterrichtseinheit. A4 thematisiert das Lernziel nicht.	0,93
II Eliciting-Evidence (EE) Ermittlung des aktuellen Lernstandes	B Typ der Frage	Die Lehrperson stellt ... B1 Eliciting-Evidence-Fragen zum erfolgten Lernen und macht damit die Denkweisen, Rechenwege und Lernprozesse einer Schülerin oder eines Schülers *retrospektiv* sichtbar. B2 kognitiv aktivierende Deep-Reasoning-Fragen und regt die Schülerinnen und Schüler dadurch zu komplexeren Denkprozessen *bei der Aufgabenbearbeitung* an.	0,81
	C Schüleräußerung	C1 Die Schülerinnen und Schüler erklären ihre Rechenwege oder mathematischen Denkprozesse.	
III Self-Assessment (SA) Schülerinnen und Schüler schätzen ihr Lernen und ihren Lernerfolg selbst ein	D Anleitung	Die Lehrperson leitet die Schülerinnen und Schüler dazu an, ... D1 den eigenen Lernprozess zu beurteilen und die nächsten Schritte des Lernens abzuleiten (Grading and What's next?). D2 den eigenen Lernprozess zu beschreiben und zu beurteilen (Describing and Grading). D3 den eigenen Lernprozess zu beurteilen (Grading).	0,89
	E Schüleraktivität	E1 Die Lehrperson moderiert das Self-Assessment. E2 Die Schülerinnen und Schüler führen das Self-Assessment selbstständig durch.	

Tab. 1 (Fortsetzung)

Strategie	Facette	Kategorien	Cohens Kappa
IV Peer-Assessment (PA) Schülerinnen und Schüler schätzen das Lernen und den Lernerfolg ihrer Peers ein	F Anleitung	Die Lehrperson leitet die Schülerinnen und Schüler dazu an, ... F1 sich gegenseitig den Lernprozess zu beschreiben und zu beurteilen. F2 sich gegenseitig den Lernprozess zu beurteilen. F3 sich gegenseitig den Lernprozess zu beschreiben.	0,95
	G Schüleraktivität	G1 Die Lehrperson moderiert das Peer-Assessment. G2 Die Schülerinnen und Schüler führen das Peer-Assessment selbstständig durch.	
V Feedback-Interaktion (FBI) Interaktionen zwischen Lehrpersonen und Schülerinnen und Schülern	H Initiierung	Die Interaktion wird initiiert durch ... H1 die Lehrperson. H2 Schülerinnen und Schüler. H3 nicht erkennbare Personen.	0,93
	I Beteiligte Schülerinnen und Schüler	Die Lehrperson interagiert mit ... I1 einer Schülerin/einem Schüler. I2 2 Schülerinnen und Schülern. I3 3–5 Schülerinnen und Schülern. I4 mehr als 5 Schülerinnen und Schülern. I5 nicht erkennbar vielen Personen.	

Tab. 2 Ratinginstrument zur Beurteilung der Qualität der in den videografierten Doppelstunden zur Einführung des halbschriftlichen Dividierens eingesetzten fünf Strategien von formativem Assessment

Strategie	Items	ICC
I Lernziel (LZ) Nennung und Besprechung des Lernziels	*LZ-1 Vorwissen:* Die Lehrperson stellt beim Vorstellen der Lernziele Bezüge zum Vorwissen der Schülerinnen und Schüler her. *LZ-2 Bedeutung:* Die Lehrperson erläutert die Bedeutung der Lernziele, indem sie einen Bezug zum Alltag, zu einer Problemstellung oder zum weiteren Wissensaufbau herstellt und die Lernziele dadurch in einen größeren Zusammenhang einordnet. *LZ-3 Diskussion/Anleitung:* Die Lehrperson entwickelt mit den Schülerinnen und Schülern ein gemeinsames Verständnis der Lernziele und stellt sicher, dass die Schülerinnen und Schüler die Lernziele verstanden haben. *LZ-4 Formulierung/Inhalt der Lernziele:* Die Lernziele beziehen sich auf mathematische Inhalte, beschreiben konkretes Verhalten und sind aus der Perspektive der Schülerinnen und Schüler formuliert. *LZ-5 Vorschau der Lehrperson auf die Erreichung der Lernziele im Unterrichtsverlauf:* Die Lehrperson erläutert zu den Lernzielen den weiteren Unterrichtsverlauf und wie die Ziele erreicht werden sollen. Lernziele und geplanter Unterrichtsablauf sind kohärent. *LZ-6 Unterrichtsabschluss:* Die Lehrperson fasst am Ende die Unterrichtsergebnisse und die Erkenntnisse bezüglich des zu Beginn explizierten Ziels oder der Problemstellung zusammen. *LZ-7 Mathematische/nicht mathematische Lernziele:* Die Lehrperson nennt mathematische/nicht mathematische Lernziele. Pro Item wurde eines von 4 Qualitätsniveaus[a] bestimmt.	0,93
II Eliciting-Evidence (EE) Ermittlung des aktuellen Lernstandes	*EE-1 Schülerbeteiligung:* Die Schülerinnen und Schüler haben Zeit und Raum, um ihr Denken im Unterricht angemessen und ausführlich zu verbalisieren. *EE-2 Anregung:* Die Lehrperson regt die Schülerinnen und Schüler durch effektive Fragestrategien dazu an, ihre Denkwege oder Handlungen zu erklären und zu begründen. *EF-3 Monitoring:* Die Lehrperson überwacht das Lernen der Schülerinnen und Schüler aufmerksam und adressiert alle Kinder. *EE-4 Produktnutzung:* Die Lehrperson sichtet und nutzt Produkte der Schülerinnen und Schüler, um einen Einblick in den aktuellen Lernstand zu erhalten. Pro Item wurde eines von 8 Qualitätsniveaus [b] bestimmt.	0,87
III Self-Assessment (SA) Schülerinnen und Schüler schätzen ihr Lernen und ihren Lernerfolg selbst ein	*SA-1 Anleitung und Durchführung des Self-Assessment/Support durch Lehrperson:* Die Lehrperson unterstützt das Self-Assessment der Schülerinnen und Schüler. Die Anleitung zum Self-Assessment wird von der Lehrperson klar verständlich formuliert bzw. baut auf bereits bekannten Unterrichtsabläufen auf. *SA-2 Anspruchsniveau*: Das Self-Assessment wirkt auf die Schülerinnen und Schüler kognitiv aktivierend. *SA-3 Nutzung für den weiteren Unterricht:* Das Self-Assessment ist in den Unterricht eingebettet und die Ergebnisse aus dem Self-Assessment werden für die nachfolgenden Lernschritte genutzt. Pro Item wurde eines von 4 Qualitätsniveaus bestimmt.	0,87

Tab. 2 (Fortsetzung)

Strategie	Items	ICC
IV Peer-Assessment (PA) Schülerinnen und Schüler schätzen das Lernen und den Lernerfolg ihrer Peers ein	*PA-1 Anleitung und Durchführung des Peer-Assessment/Support durch Lehrperson:* Die Lehrperson unterstützt das Peer-Assessment der Schülerinnen und Schüler. Die Anleitung zum Peer-Assessment wird von der Lehrperson klar verständlich formuliert bzw. baut auf bereits bekannten Unterrichtsabläufen auf. *PA-2 Anspruchsniveau:* Die Schülerinnen und Schüler reflektieren gegenseitig die Qualität von Lösungswegen und Ergebnissen, welche kognitiv anspruchsvollen Tätigkeiten wie Vergleichen, Analysieren, Beurteilen etc. verlangen. *PA-3 Nutzung für den weiteren Unterricht:* Das Peer-Assessment ist in den Unterricht eingebettet und die Ergebnisse aus dem Peer-Assessment werden für die nachfolgenden Lernschritte genutzt. Pro Item wurde eines von 4 Qualitätsniveaus bestimmt.	0,93
V Feedback-Interaktion (FBI) Lehrperson teilt in Interaktionen den Schülerinnen und Schülern ihre Einschätzung zu ihren Lernprozessen und Lernerfolgen mit	*FBI-1 Feed Up:* Das Feedback verdeutlicht Sinn oder Ziele von Aufgaben und Tätigkeiten im Unterricht und ordnet diese in einen größeren Kontext ein. *FBI-2 Feed Back:* Das Feedback regt die Beschreibung und die Evaluation des Lernstands und/oder des Lernprozesses an. *FBI-3 Feed Forward:* Das Feedback erweitert ausgehend vom aktuellen Stand das Verstehen und die Handlungen der Schülerinnen und Schüler. *FBI-4 Fokus:* Das Feedback ist spezifisch, auf mathematische Inhalte bezogen und fokussiert auf die Aufgabenbearbeitung und den Prozess des Lernens. *FBI-5 Aktivierung:* Das Feedback unterstützt das selbstständige Denken, die Selbstregulation und die Metakognition. *FBI-6 Adaptivität* (subjektbezogene und situationsspezifische Passung): Das Feedback ist an den Kenntnisstand der Schülerinnen und Schüler angepasst. *FBI-7 Motivation & Emotion:* Die Lehrpersonen unterstützt mit ihrem Feedback die Motivation und das Engagement der Schülerinnen und Schüler. *FBI-8 Wechselseitigkeit:* Es bestehen lernunterstützende wechselseitige Gesprächsbeiträge zwischen der Lehrperson und den Schülerinnen und Schülern. Pro Item wurde eines von 8 Qualitätsniveaus bestimmt.	0,85

[a]Beispiel für eine vierstufige Ratingskala im Anhang B
[b]Beispiel für eine achtstufige Ratingskala im Anhang A

strukturiertes Beobachtungsinstrument erarbeitet, beginnend mit grob beschreibenden niedriginferenten Kodierungen, das heißt mit Basiskodierungen, die sich ohne viel Interpretationsspielraum auf direkt beobachtbare Merkmale beziehen (Hugener 2006). Das Ziel der Basiskodierungen bestand darin, einen Überblick über die gefilmten Doppelstunden zu erstellen und die Analyseeinheiten für die anschließenden Kodier- und Ratingdurchgänge festzulegen. Als Basiskodierung wurden (in Anlehnung an Klieme et al. 2006) (a) der Unterrichtsstatus, (b) der Unterrichtsinhalt und (c) die Sozialform erfasst. Im ersten Durchgang für den *Unterrichtsstatus* wurde mit zwei disjunkten Kategorien (Unterricht/kein Unterricht) die effektive Unterrichtszeit bestimmt. Anschließend wurde innerhalb der effektiven Unterrichtszeit zwischen mathematikbezogenem und nicht mathematikbezogenem *Unterrichtsinhalt* unterschieden (= mathematische Unterrichtszeit). Im letzten Durchgang wurde

schließlich die *Sozialform* bestimmt, das heißt, es wurde kodiert, wie lange die Schülerinnen und Schüler selbstständig in Schülerarbeitsphasen gearbeitet hatten und wie viel Zeit auf den von der Lehrperson angeleiteten Klassenunterricht entfallen war (insgesamt 11 disjunkte Kategorien). Die auf diese Weise ermittelte mathematische Unterrichtszeit bildete die Analyseeinheit für die nachfolgende Kodierung und das Rating des formativen Assessments.

Die Basiskodierungen wurden von zwei unabhängigen Kodierenden hinsichtlich der Interrater-Reliabilität geprüft. Die AC1-Werte zur Bestimmung der Interrater-Reliabilität liegen zwischen 0,92 und 0,98 und können als sehr hoch bezeichnet werden. Die Qualität der Datenaufbereitung konnte außerdem durch das Vorgehen in mehreren Kodierdurchgängen erhöht werden, da nicht übereinstimmende Kodierungen aus dem jeweils vorhergehenden Durchgang beim darauf aufbauenden Durchgang laufend konsensual korrigiert wurden. Auf der Grundlage dieser niedriginferenten Basiskodierungen wurden die mittelinferenten Kodierungen für die Häufigkeit und die Dauer des formativen Assessments und anschließend die hochinferenten Ratings zur Bestimmung der Qualität bei der Anwendung der fünf Strategien im laufenden Unterricht entwickelt. Für alle Instrumente wurde ein Manual mit präzisen Beschreibungen des Beobachtungsgegenstands und Handlungsanweisungen, ergänzt mit Ankerbeispielen und Erläuterungen für den Umgang mit Zweifelsfällen, erstellt.

Die Kodierungen und Ratings wurden von insgesamt sieben Projektmitarbeitenden durchgeführt, wobei in der Regel arbeitsteilig vorgegangen wurde. Alle Kodierenden und Ratenden hatten zuvor an einem mehrtägigen Training teilgenommen, in dessen Rahmen sie sich mit dem Manual vertraut machten und die Instrumente probeweise anwenden konnten. Nach selbstständigen Probekodierungen und Proberatings folgten jeweils intensive Diskussionen in der Gruppe. Anschließend wurden notwendige Anpassungen und Präzisierungen in den Manualen vorgenommen. Erst nach erfolgreichen Reliabilitätsprüfungen wurden die Kodierungen und Ratings durchgeführt.

4.4 Kodierungen zur Erfassung der Häufigkeit und der Dauer

In seiner inhaltlichen Struktur bildet das Kodierinstrument die in Abschn. 2 theoretisch beschriebenen fünf Strategien des formativen Assessments nach Black und Wiliam (2009) ab. Auf dieser Basis wurden die Strategien, Facetten und die meisten der Kategorien mit Blick auf eine informelle Anwendung des formativen Assessments im laufenden Unterricht deduktiv festgelegt. Die Ausdifferenzierung der ein bis maximal fünf disjunkten Kategorien erfolgte anschließend induktiv am Material (Tab. 1).

4.4.1 Anwendung des Kodierinstruments

Für die Bestimmung der zeitlichen Einheiten, denen die Kodes zugewiesen werden, kann grundsätzlich zwischen Ereignis- und Zeitstichprobe unterschieden werden (Event-Sampling versus Time-Sampling). Bei Ereignisstichproben bestimmt das zu erfassende Ereignis den Anfangs- und den Endpunkt für die Kodesetzung. Bei Zeitstichproben hingegen wird nach einem festgelegten Zeitintervall (z. B. 10 Sek.)

kodiert, ob das interessierende Verhalten oder Merkmal beobachtbar ist (Pauli 2012). In der vorliegenden Teilstudie wurden beide Verfahren verwendet. Zeitstichproben sind ressourcenschonender, da die Bestimmung der Anfangs- und Endpunkte wegfällt und sich das Kodieren an festen Zeitintervallen orientiert (Berner et al. 2013). Deshalb wurde hauptsächlich mit Zeitstichproben gearbeitet, während Ereignisstichproben dann die Basis bildeten, wenn die inhaltliche Bestimmung des Anfangs- und Endpunkts für eine weiterführende vertiefende Datenaufbereitung von Bedeutung war.

Die Strategien *Lernziel, Eliciting-Evidence, Self-Assessment* und *Peer-Assessment* wurden mit der Software Videograph (Version 4.2.1.1.X3, Rimmele 2004) als Zeitstichproben mit Intervallen von 10 s kodiert. Pro Intervall wurde jeweils nur eine Kategorie einer Facette vergeben, und zwar jene, die zeitlich am längsten gedauert hatte. Im Gegensatz dazu wurde die Strategie *Feedback-Interaktion* als Ereignisstichprobe mit MAXQDA (Version 2018) erfasst, damit verkettete Gesprächsbeiträge durch eine später erfolgende vertiefende Interaktionsanalyse ausgewertet werden konnten (Ruelmann, in Vorb.). Während die anderen vier Strategien sowohl für den Klassenunterricht als auch für die Schülerarbeitsphasen der mathematischen Unterrichtszeit kodiert worden waren, wurde die Feedback-Interaktion nur für die Schülerarbeitsphasen der Doppelstunde kodiert, da die Lehrpersonen in solchen Phasen einzelne oder mehrere Schülerinnen und Schüler in dialogischen Interaktionen gezielt durch Feedbacks individuell in ihrem Lernprozess unterstützt hatten.

Rund 10 % der Videoaufnahmen wurden zunächst von je zwei Personen der Forschungsgruppe unabhängig voneinander kodiert. Nach dem Erreichen einer genügend hohen Interrater-Reliabilität (vgl. die Werte für Cohens Kappa in Tab. 1) wurden die weiteren Unterrichtsvideos von beiden Mitarbeitenden einzeln kodiert. Das Vorgehen bei der Kodierung wird nachfolgend am Beispiel der Strategie *Lernziel* und Facette A illustriert. Diese Facette umfasst die vier Kategorien A1 Lernziel begründen, A2 Lernziel einordnen, A3 Lernziel besprechen und A4 Art der Thematisierung nicht erkennbar. Pro 10-Sekunden-Intervall wurde bestimmt, ob die Lehrperson das Lernziel begründet (A1), eingeordnet (A2), besprochen (A3) oder es nicht thematisiert (A4) hatte oder ob eine andere Facette mit ihren Kategorien zu kodieren war. Wie bereits erwähnt wurden im Kodiermanual zu jeder Kategorie Erläuterungen, Erklärungen, Regeln (z. B. für den Umgang mit Zweifelsfällen) und Ankerbeispiele festgehalten.

4.4.2 Auswertung

Vor der Datenauswertung wurde die Einhaltung der im Kodiermanual festgelegten Bedingungen und Verbindungen zwischen den einzelnen Durchgängen der Basiskodierung und den Kodierungen der fünf Strategien des formativen Assessments überprüft. Inkonsistenzen wurden gegebenenfalls bereinigt. Für die Auswertung der Kodierungen wurden mit den Statistikprogramm IBM SPSS (Version 25) deskriptive Statistiken erstellt. Um die Gesamtdauer zu ermitteln, wurden die für jede Lehrperson unter der jeweiligen Kategorie, Facette bzw. Strategie kodierten 10-Sekunden-Intervalle addiert. Die Häufigkeiten wurden als prozentuale Anteile der jeweiligen Strategie an der mathematischen Unterrichtszeit bestimmt.

4.5 Ratings zur Beurteilung der Qualität

Auch das Ratinginstrument bildete die fünf Strategien nach Black und Wiliam (1998) in ihrer informellen Umsetzung im alltäglichen Unterricht ab. Die Qualitätsabstufungen orientierten sich an den Kernmerkmalen der Qualität des formativen Assessments (vgl. Abschn. 2). Wie Tab. 2 zu entnehmen ist, wurden zu jeder der fünf Strategien Items und für jedes Item Qualitätsniveaus formuliert.

4.5.1 Anwendung des Ratinginstruments

Das Rating zu den Strategien *Lernziel, Self-Assessment* und *Peer-Assessment* erfolgte für sämtliche Stellen innerhalb der mathematischen Unterrichtszeit, an denen das Vorkommen der jeweiligen Strategie vorgängig kodiert worden war. Weil sich bei der Kodierung ergeben hatte, dass die Strategien *Eliciting-Evidence* und *Feedback-Interaktion* fast durchgängig über die gesamte mathematische Unterrichtszeit hinweg beobachtbar waren, wurde das Rating dieser beiden Strategien für je eine zusammenhängende 15-minütige Sequenz aus dem Klassenunterricht und aus der Schülerarbeitsphase durchgeführt (insgesamt 30 min pro Video). Es wurden bevorzugt Videoabschnitte geratet, in welchen mindestens 15 min am Stück Klassenunterricht oder Schülerarbeitsphase vorkamen. Gab es innerhalb eines Videos mehrere solcher Stellen, wurden die ersten 15 min hiervon für das Rating verwendet. Fanden sich in einem Video weniger oder nicht zusammenhängende 15 min Klassenunterricht oder Schülerarbeitsphase, begann das Rating beim ersten Videoausschnitt mit Klassenunterricht oder Schülerarbeit und zog sich über mehrere Abschnitte hinweg.

Die Qualität der Anwendung der Strategien *Eliciting-Evidence* und *Feedback-Interaktion* wurde anhand von acht Niveauabstufungen beurteilt.[2] Die Anwendung der Ratingskala wurde aus dem Classroom Assessment Scoring System von Pianta et al. (2012) übernommen, welches ebenfalls für 15-minütige Videoausschnitte konzipiert worden war. Im Ratingmanual wurden für jedes Item die Kriterien für eine geringe, eine mittlere und eine hohe Qualitätsausprägung definiert (siehe Anhang A). Die Vergabe der Werte auf der achtstufigen Skala erfolgte, indem bestimmt wurde, ob eine geringe, mittlere oder hohe Qualitätsausprägung gemäß Manual vorlag. Bei hoher Ausprägung, das heißt, wenn alle Qualitätsmerkmale erfüllt waren, wurde der Höchstwert 7 vergeben. Der Wert 6 wurde dann gesetzt, wenn die Qualitätsmerkmale nicht vollständig erfüllt waren und noch ein oder zwei Indikatoren der mittleren Qualitätsausprägung zu beobachten waren. Die Vergabe des Werts 5 erfolgte, wenn die Merkmale der mittleren Qualitätsausprägung vorhanden und zusätzlich ein oder zwei Indikatoren der hohen Ausprägung feststellbar waren. Trafen die Merkmale der mittleren Ausprägung vollständig zu, lag der Wert 4 vor. Waren diese Merkmale hingegen unvollständig erfüllt, das heißt, wenn ein oder zwei Indikatoren der geringen Qualitätsausprägung vorhanden waren, wurde der Wert 3 zugeordnet. Lagen die Merkmale der geringen Ausprägung und ein oder zwei Indikatoren der mittleren Ausprägung vor, wurde der Wert 2 vergeben. Der Wert 1 wurde gewählt, wenn vorwiegend die Merkmale der geringen Ausprägung vorhanden waren. Der Wert 0 lag

[2] 0 = Item nicht einschätzbar; 1, 2 = geringe Qualität; 3, 4, 5 = mittlere Qualität; 6, 7 = hohe Qualität.

vor, wenn die jeweilige Strategie im Unterricht vorkam und entsprechende Kodes gesetzt wurden, der Videoabschnitt jedoch zu kurz war, um das jeweilige Item zu raten (vgl. Beispiel in Anhang A). Falls die gesamte Strategie des formativen Assessments nicht beobachtet werden konnte, wurde kein Wert für das Rating vergeben (Missing).

Gemäß allgemeinen inhaltsanalytischen Standards sind Skalenabstufungen den Differenzierungsmöglichkeiten des Materials anzupassen (Schreier 2014). Diesem Grundsatz entsprechend schien der Einsatz einer achtstufigen Ratingskala für zeitlich vergleichsweise umfangreiches Videomaterial von ca. 15 min in der vorliegenden Studie adäquat zu sein, nicht jedoch für die Strategien *Lernziel, Self-Assessment* und *Peer-Assessment* mit einer durchschnittlichen Dauer von 0,84 bis 8,19 min (Tab. 3). Aus diesem Grund wurde die Qualität bei diesen drei Strategien anhand von vierstufigen Skalen eingeschätzt.[3] Auch hier wurde für jedes Item eine hohe, eine mittlere und eine geringe Ausprägung definiert. Bei der Strategie *Lernziel* beispielsweise reichte das Spektrum des Items LZ-1 zum Vorwissen von einer reichhaltigen und aktiven Verknüpfung von Vorwissen und Lernziel über eine schwache und wenig ausgeprägte bis hin zu einer ausbleibenden aktiven Verknüpfung (vgl. Beispiel in Anhang B). Der Wert 0 wurde hier vergeben, wenn die jeweilige Strategie im Unterricht zwar vorkam, für die Beurteilung des einzelnen Items jedoch zu wenig Videomaterial vorlag.

Nach erfolgter Schulung anhand von rund 10 % der Videoaufnahmen wurden die Qualitätsausprägungen von zwei Mitarbeitenden der Forschungsgruppe in den weiteren Unterrichtsvideos für jede der fünf Strategien unabhängig voneinander eingeschätzt (vgl. den ICC-Wert in Tab. 2).

4.5.2 Auswertung

Zur Überprüfung der Ratingskalen wurde trotz kleiner Stichprobengröße eine explorative Faktorenanalyse durchgeführt (Hauptkomponentenanalyse mit obliquer Rotation). Diese ergab, dass sich die Items der Strategien *Lernziel, Self-Assessment* und *Peer-Assessment* erwartungsgemäß verteilten (standardisierte Faktorladungen zwischen 0,49 und 0,92); einzig das Lernziel-Item LZ-6 musste ausgeschlossen werden, weil es nicht mit den anderen Items korrelierte. Die Items der Strategien *Eliciting-Evidence* und *Feedback-Interaktion* hingegen luden jeweils zusammen auf je einen Faktor für die Schülerarbeitsphase und einen Faktor für den Klassenunterricht. Die weitere Überprüfung der Zusammenhänge zwischen *Eliciting-Evidence* und *Feedback-Interaktion* zeigte, dass die beiden zugrunde gelegten Ratingskalen untrennbar hoch korrelierten (Klassenunterricht: $r = 0{,}92$, $p < 0{,}001$; Schülerarbeitsphasen: $r = 0{,}91$ $p < 0{,}001$). Allerdings hingen die beiden Strategien über Klassenunterricht und Schülerarbeitsphase hinweg deutlich tiefer zusammen (*Eliciting-Evidence:* $r = 0{,}64$, $p < 0{,}001$; *Feedback-Interaktion:* $r = 0{,}71$, $p < 0{,}001$), so dass sie nicht über die Unterrichtsphasen hinweg zusammengefasst werden konnten. Deshalb werden die Ratingwerte zu diesen beiden Strategien des formativen Assessments in

[3] 0 = Item nicht einschätzbar; 1 = geringe Qualität; 2 = mittlere Qualität; 3 = hohe Qualität.

Tab. 3 Kodierung der unterrichtlichen Sichtstruktur: Häufigkeit der Strategien des formativen Assessments nach Strategie, Facette und Kategorie (*M*, *SD*, Min./Max.)

Strategie (I–V)/Facette (A–I)/Kategorie	Anteil Lehrpersonen, die FA nutzen	Dauer in Minuten		Prozentualer Anteil der mathematischen Unterrichtszeit[b]	
	Anzahl (%)[a]	*M* (*SD*)	Min./Max	*M* (*SD*)	Min./Max
I Lernziel (LZ)	*45 (87)*	*1,40 (1,78)*	*0–8,50*	*2 (2)*	*0–10*
A Lernziel	*45 (87)*	*1,40 (1,78)*	*0–8,50*	*2 (2)*	*0–10*
A1 Lernziel begründen	3 (6)	0,15 (0,73)	0–4,83	0 (1)	0–6
A2 Lernziel einordnen	27 (52)	0,49 (1,11)	0–7,33	1 (1)	0–9
A3 Lernziel besprechen	43 (83)	0,76 (0,90)	0–4,67	1 (1)	0–5
A4 Nicht erkennbar thematisiert	1 (2)	0,003 (0,02)	0–0,17	0 (0)	0–0
II Eliciting Evidence (EF)	*52 (100)*	*4,85 (3,13)*	*0,33–12,17*	*6 (4)*	*1–15*
B Typ der Frage	*51 (98)*	*2,02 (1,24)*	*0–5,33*	*2 (2)*	*0–7*
B1 Eliciting-Evidence-Frage	48 (92)	1,40 (1,10)	0–5,17	2 (1)	0–6
B2 Deep-Reasoning-Frage	46 (88)	0,62 (0,62)	0–3,67	1 (1)	0–4
C Schüleräußerung (Aktivität)	*51 (98)*	*3,10 (2,41)*	*0–10,50*	*4 (3)*	*0–12*
C1 EE/Deep-Reasoning	51 (98)	3,10 (2,41)	0–10,50	4 (3)	0–12
III Self-Assessment (SA)	*40 (77)*	*0,84 (0,90)*	*0–3,33*	*1 (1)*	*0–4*
D Anleitung	*37 (71)*	*0,60 (0,60)*	*0–2,33*	*1 (1)*	*0–4*
D1 Grading and What's next?	28 (54)	0,32 (0,44)	0–2,33	0 (1)	0–4
D2 Describing and Grading	8 (15)	0,05 (0,12)	0–0,50	0 (0)	0–1
D3 Grading	16 (31)	0,23 (0,44)	0–1,67	0 (1)	0–2
E Aktivität	*18 (35)*	*0,25 (0,54)*	*0–2,33*	*0 (1)*	*0–3*
E1 Von Lehrperson geleitet	15 (29)	0,22 (0,51)	0–2,17	0 (1)	0–3
E2 Von Schülerin/Schüler selbstständig durchgeführt	5 (10)	0,03 (0,11)	0–0,67	0 (0)	0–1
IV Peer-Assessment (PA)	*46 (88)*	*8,19 (7,39)*	*0–33,17*	*10 (9)*	*0–39*
F Anleitung	*45 (87)*	*0,93 (0,91)*	*0–4,17*	*1 (1)*	*0–5*
F1 Describing and Grading	6 (12)	0,13 (0,46)	0–2,50	0 (1)	0–3
F2 Grading	2 (4)	0,01 (0,05)	0–0,33	0 (0)	0–0
F3 Describing	45 (87)	0,79 (0,65)	0–2,33	1 (1)	0–3

Tab. 3 (Fortsetzung)

Strategie (I–V)/Facette (A–I)/Kategorie	Anteil Lehrpersonen, die FA nutzen	Dauer in Minuten		Prozentualer Anteil der mathematischen Unterrichtszeit[b]	
	Anzahl (%)[a]	*M* (*SD*)	Min./Max	*M* (*SD*)	Min./Max
G Aktivität	*46 (88)*	*7,27 (6,67)*	*0–29,17*	*9 (8)*	*0–37*
G1 Von Lehrperson geleitet	43 (83)	5,20 (4,71)	0–17,83	6 (6)	0–23
G2 Von Schülerin/Schüler selbstständig durchgeführt	18 (35)	2,07 (3,82)	0–14,83	3 (5)	0–18
V Feedback-Interaktion (FBI)[c]	*52 (100)*	*33,76 (9,01)*	*14,60–51,10*	*42 (11)*	*17–61*
H Initiierung der Interaktion	*52 (100)*	*33,76 (9,01)*	*14,60–51,10*	*42 (11)*	*17–61*
H1 Von Lehrperson initiiert	52 (100)	16,53 (8,99)	0,73–35,36	20 (11)	1–44
H2 Von Schülerin/Schüler initiiert	51 (98)	16,34 (9,48)	0–42,93	20 (12)	0–54
H3 Nicht erkennbar	30 (58)	0,88 (1,53)	0–7,05	1 (2)	0–8
I Anzahl der beteiligten Schülerinnen und Schüler	*52 (100)*	*33,76 (9,01)*	*14,60–51,10*	*42 (11)*	*17–61*
I1 1 Schülerin/Schüler	52 (100)	23,97 (10,71)	2,45–44,85	29 (13)	3–56
I2 2 Schülerinnen und Schüler	42 (81)	5,17 (6,66)	0–30,60	6 (8)	0–34
I3 3–5 Schülerinnen und Schüler	38 (73)	4,25 (5,41)	0–21,09	5 (7)	0–26
I4 >5 Schülerinnen und Schüler	4 (8)	0,11 (0,51)	0–3,09	0 (1)	0–4
I5 Nicht erkennbar	14 (27)	0,24 (0,66)	0–3,62	0 (1)	0–4

$n = 52$

[a]Alle Prozentwerte sind auf ganze Zahlen gerundet

[b]100 % = mathematische Unterrichtszeit; $M = 81{,}58$ min; $SD = 6{,}41$; Range: 60,66–92,00 min (Ausschluss von nicht mathematischem Unterricht und keinem Unterricht)

[c]Die Gesamtzeiten für die Facetten H „Initiierung" und I „Anzahl der beteiligten Schülerinnen und Schüler" sind dieselben, weil jeweils einer der Codes H1–H3 bzw. I1–I5 für jedes als Feedback-Interaktion definierte Intervall vergeben wurde

Tab. 4 separat berichtet, einmal für die Schülerarbeitsphase und einmal für den Klassenunterricht.

Zur Beurteilung der internen Konsistenz wurde der omega Wert berechnet. Wie in Tab. 4 aufgeführt, bewegten sich die Werte zwischen 0,84 und 0,99 und können als hoch bis ausgezeichnet bezeichnet werden. Eingesetzt wurden die Statistikprogramme IBM SPSS (Version 25) und R (Version 3.6.0, Paket „psych").

Für die folgenden Analysen wurde aus den gerateten Items für jede Strategie ein Mittelwert berechnet.

Tab. 4 Eingeschätzte durchschnittliche Qualität (*M*, *SD*, Range) der Assessment-Strategien (linke Tabellenhälfte), Zusammenhänge zwischen den Assessment-Strategien (rechte Tabellenhälfte) und omega Werte zur internen Konsistenz (gefettete Werte auf der Diagonale) der Ratingskalen

Strategie	Qualität der Strategieanwendung[a]			Zusammenhänge zwischen den Strategien				
	n	*M* (*SD*)	Range	LZ	SA	PA	EE-FBI SA	EE-FBI KU
Lernziel (LZ)	45	1,05 (0,64)	0–3,00	**(0,89)**	–	–	–	–
Self-Assessment (SA)	38[b]	1,14 (0,53)	0,33–2,67	–0,1	**(0,84)**	–	–	–
Peer-Assessment (PA)	46	1,70 (0,74)	0,67–3,00	0,002	0,42*	**(0,92)**	–	–
Eliciting-Evidence und Feedback-Interaktion in Schülerarbeitsphase (EE-FBI SA)	52	3,19 (1,01)	1,75–6,08	0,10	0,01	0,62**	**(0,98)**	–
Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht (EE-FBI KU)	50	3,45 (1,07)	1,58–5,67	–0,03	0,14	0,67**	0,70**	**(0,99)**

*$p < 0{,}05$, **$p < 0{,}01$, Pearson Korrelation, zweiseitig

[a]Einschätzung der Strategien LZ, SA und PA als Mittelwert über 3 bis 7 Items mit vierstufiger (0–3) und der Strategien EE und FBI über 4 bis 8 Items mit achtstufiger Skala (0–7)

[b]Bei zwei Lehrpersonen waren die Videoausschnitte zum Self-Assessment zu kurz (weniger als 10 s), um ein Rating vornehmen zu können

5 Ergebnisse

5.1 Häufigkeit und Dauer der Strategien des formativen Assessments (Kodierung)

Mit Blick auf Forschungsfrage 1 lässt sich Tab. 3 zur *Häufigkeit* der Strategieanwendung entnehmen, dass die große Mehrzahl der Lehrpersonen jede der fünf Strategien nutzt. Die Strategien *Eliciting-Evidence* und *Feedback-Interaktion* werden von allen 52 Lehrpersonen der Stichprobe eingesetzt (100 %). Demgegenüber lassen sich mit abnehmender Häufigkeit die Strategie *Peer-Assessment* bei 46 Lehrpersonen (88 %), die Strategie *Lernziel* bei 45 Lehrpersonen (87 %) und die Strategie *Self-Assessment* bei 40 Lehrpersonen (77 %) feststellen.

Die *Dauer* des Einsatzes der Strategien insgesamt wie auch der jeweils zugehörigen Facetten und Kategorien wird in Tab. 3 zum einen *in Minuten* und zum anderen als *prozentualer Anteil* der mathematischen Unterrichtszeit von durchschnittlich 81,6 min angegeben. Die Ergebnisse zeigen, dass die Strategie *Feedback-Interaktion* durchschnittlich während 33,8 min und somit mit einem 42 %-Anteil an der mathematischen Unterrichtszeit mit Abstand am längsten beobachtbar ist. Feedback-Interaktionen mit einer einzelnen Schülerin oder einem einzelnen Schüler kommen am häufigsten vor (29 % der mathematischen Unterrichtszeit) und sind etwa gleich häufig von der Lehrperson und den betreffenden Schülerinnen und Schülern initi-

iert. Die Strategie *Peer-Assessment* kommt mit durchschnittlich 8,2 min und einem Anteil von 10 % an der mathematischen Unterrichtszeit am zweithäufigsten vor. Das von der Lehrperson geleitete Peer-Assessment macht hier den größten Anteil aus (5,2 min; 6 %). Mit abnehmendem zeitlichem Anteil folgen die Strategien *Eliciting-Evidence* (4,9 min; 6 %), *Lernziel* (1,4 min; 2 %) und *Self-Assessment* (0,8 min; 1 %). Bei der Strategie *Eliciting-Evidence* lässt sich feststellen, dass retrospektive Eliciting-Evidence-Fragen (1,4 min; 2 %) häufiger gestellt werden als Deep-Reasoning-Fragen zur aktuellen Aufgabenbearbeitung (0,6 min; 1 %).

5.2 Qualität der Strategien des formativen Assessments (Rating)

Auf Forschungsfrage 2 Bezug nehmend zeigen die Ergebnisse zum Rating in Tab. 4, dass die durchschnittlichen Qualitätsbeurteilungen der Strategien *Lernziel* ($M = 1{,}05$), *Self-Assessment* ($M = 1{,}14$) sowie *Peer-Assessment* ($M = 1{,}70$) im unteren Bereich der vierstufigen Skala liegen. Eine eher geringe Qualitätsausprägung bedeutet beispielsweise bei der Strategie *Lernziel*, dass eine Lehrperson die Lernziele oder das Unterrichtsthema nur benennt, das heißt oberflächlich thematisiert und bei der Planung, Gestaltung und Beurteilung von Lernprozessen kaum darauf verweist. Die geringe Qualitätsausprägung bei den Strategien *Self-Assessment* und *Peer-Assessment* weist darauf hin, dass das Verstehen von Lernprozessen kaum fokussiert wird, keine Begründungen für die gewählten Lösungswege verlangt werden und nur selten die Gelegenheit besteht, über den gewählten Lösungsweg zu sprechen und sich über unterschiedliche Lösungswege auszutauschen.

Die Ratingwerte für *Eliciting-Evidence* und *Feedback-Interaktion* werden in den Tab. 4 und 5 zusammengefasst und differenziert nach Schülerarbeitsphase und Klassenunterricht unter *Eliciting-Evidence und Feedback-Interaktion in Schülerarbeitsphase* und *Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht* berichtet. Die entsprechenden Werte liegen auf der achtstufigen Skala knapp unter der mittleren Qualitätsausprägung (3,19 bzw. 3,45). Dies bedeutet, dass der Sinn und die Ziele von Aufgaben und Tätigkeiten vage einordnet werden (Feed Up), der Lernstand manchmal beschrieben, erfasst und evaluiert wird (Feed Back) und das

Tab. 5 Zusammenhänge zwischen Häufigkeit und Qualität der Assessment-Strategien

Qualität	*n*	Häufigkeit[a]			
		Lernziel (LZ)	Self-Assessment (SA)	Peer-Assessment (PA)	Feedback-Interaktion (FBI)
Lernziel (LZ)	45	0,46*	–	–	–
Self-Assessment (SA)	38	–	0,59*	–	–
Peer-Assessment (PA)	46	–	–	0,80*	–
Eliciting-Evidence und Feedback-Interaktion in Schülerarbeitsphase[b] (EE-FBI SA)	52	–	–	–	0,39*

* $p < 0{,}01$, Pearson Korrelation zweiseitig
[a]Prozentualer Anteil an mathematischem Unterricht
[b]Qualitätswerte von Eliciting-Evidence und Feedback können nicht voneinander getrennt werden

Verstehen und die Handlungen teilweise durch Impulse erweitert werden (Feed Forward). Die mit der Strategie *Feedback-Interaktion* verbundenen Fragen der Strategie *Eliciting-Evidence* ermöglichen es einigen Schülerinnen und Schülern, eigene Gedanken und Überlegungen einzubringen. Zudem ziehen die Lehrpersonen manchmal die Produkte der Schülerinnen und Schüler bei, um Einblick in ihren Lernstand zu erhalten.

Außerdem wurde geprüft, ob sich die Qualitätsratings der verschiedenen Assessment-Strategien aufeinander beziehen. Tab. 4 zeigt, dass die Qualitätsratings untereinander teilweise korrelieren. Vergleichsweise hohe Korrelation zeigen sich zwischen *Eliciting-Evidence und Feedback-Interaktion in der Schülerarbeitsphase und Peer-Assessment* ($r = 0{,}62$, $p < 0{,}01$) sowie jene zwischen *Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht* und *Peer-Assessment* ($r = 0{,}67$, $p < 0{,}01$) und jene zwischen *Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht* und *Eliciting-Evidence und Feedback-Interaktion in der Schülerarbeitsphase* ($r = 0{,}70$, $p < 0{,}01$). Zudem zeigt sich eine Korrelation zwischen *Peer-* und *Self-Assessment* ($r = 0{,}42$, $p \leq 0{,}05$). Die Kennwerte zur internen Konsistenz der Strategien sind in Tab. 4 aufgeführt (gefettete Werte auf der Diagonale).

5.3 Zusammenhang von Dauer und Qualität

Zur Klärung von Forschungsfrage 3 wurde überprüft, ob die Dauer und die Qualitätsausprägung der verschiedenen Strategien des formativen Assessments zusammenhängen. Zu diesem Zweck wurden die Ratingwerte mit der Dauer der fünf Strategien (aus Tab. 3) korreliert. Signifikant positive und hohe Korrelationen ergaben sich bei den Strategien *Lernziel* ($r = 0{,}46$, $p < 0{,}01$), *Self-Assessment* ($r = 0{,}59$, $p < 0{,}01$), *Peer-Assessment* ($r = 0{,}80$, $p < 0{,}01$) und bei *Eliciting-Evidence und Feedback-Interaktion in der Schülerarbeitsphase* ($r = 0{,}39$, $p < 0{,}01$) (Tab. 5). Zusammenfassend lässt sich somit festhalten, dass der Zusammenhang zwischen Dauer und Qualität der Strategien hoch ausfällt.

6 Diskussion

Bislang liegen sehr wenige Studien vor, in denen die Praxis des informellen formativen Assessments von unabhängigen Beobachterinnen und Beobachtern erfasst und eingeschätzt wurde. Dies gilt insbesondere für den deutschsprachigen Raum. Bei den wenigen Videostudien aus dem angloamerikanischen Raum stellt sich die Frage der Übertragbarkeit der Ergebnisse auf die Schulpraxis im deutschsprachigen Kontext (Stigler und Hiebert 1999). Zudem enthalten sie keine Aussagen zur Häufigkeit *und* zur Qualität des formativen Assessments. Angesichts dieser unzureichenden Forschungslage leisten die Befunde der berichteten Teilstudie des Forschungsprojekts TUFA zur Häufigkeit, Dauer und Qualität des informellen formativen Assessments im alltäglichen Mathematikunterricht einen ersten Beitrag zur Bearbeitung dieser Forschungslücke.

6.1 Forschungsfrage 1: Häufigkeit und Dauer

Der überwiegende Teil der teilnehmenden Lehrpersonen nutzt alle fünf erhobenen Strategien des formativen Assessments. Während die Strategien *Eliciting-Evidence* und *Feedback-Interaktion* von allen 52 Lehrpersonen der Stichprobe eingesetzt werden, kommen die Strategien *Lernziel* (45 Lehrpersonen), *Self-Assessment* (40 Lehrpersonen) und *Peer-Assessment* (46 Lehrpersonen) weniger oft zum Einsatz. Insgesamt betrachtet bestätigen die Ergebnisse weitgehend die Befunde von Altmann et al. (2010), Bürgermeister (2014) und Schmidt (2020), obschon ein Vergleich aufgrund der unterschiedlichen Operationalisierungen der Häufigkeit sowie des methodischen Vorgehens (Selbstauskünfte vs. Urteile von Beobachterinnen und Beobachtern oder Expertinnen und Experten) nur bedingt möglich ist. Abweichend ausgefallen sind bei differenzierterer Betrachtung allerdings die Befunde zu den Strategien *Self-Assessment* und *Peer-Assessment*. In der vorliegenden Teilstudie kann der Anteil derjenigen Lehrpersonen, die diese beiden Strategien einsetzen, mit 77 % bzw. 88 % als sehr hoch bezeichnet werden. In der Studie von Bürgermeister (2014) hatten die Befragten demgegenüber angegeben, diese beiden Strategien „nie“ bis „manchmal“ anzuwenden. Auch Schmidt (2020) sowie Cheng und Wang (2007) hielten fest, dass Peer- und Self-Assessments im Unterricht der befragten Lehrpersonen im Vergleich zu anderen Strategien eine relativ geringe Rolle spielen.

Feedback-Interaktionen machen mit einer zeitlichen Dauer von durchschnittlich 33,8 min einen erheblichen Anteil der im Durchschnitt 81,6 min dauernden mathematischen Unterrichtszeit aus. Es handelt sich um diejenige Strategie, die im Durchschnitt deutlich länger eingesetzt wird als jede der anderen Strategien. Es folgen die Strategien *Peer-Assessment* (8,2 min), *Eliciting-Evidence* (4,9 min) und *Lernziel* (1,4 min). Am seltensten kommt die Strategie *Self-Assessment* vor (0,8 min). Insgesamt muss in Anbetracht der Ergebnisse konstatiert werden, dass die durchschnittliche Einsatzdauer der Strategien mit Ausnahme der Feedback-Interaktionen kurz ausfällt und die Streubereiche erheblich sind.

Eine Einordnung der berichteten Ergebnisse in bestehende Forschung lässt sich nur grob vornehmen, da in den wenigen Videostudien zum formativen Assessment die Dauer nicht erfasst wurde. Es können jedoch Bezüge zur Videostudie von Krammer (2009) hergestellt werden, bei der ähnliche Themen untersucht wurden. Gemäß den Befunden von Krammer (2009) beträgt der zeitliche Anteil der individuellen Lernunterstützung in den von ihr analysierten Mathematiklektionen rund 23 % der Unterrichtszeit. Die individuelle Lernunterunterstützung weist eine konzeptionelle Nähe zu den in der vorliegenden Teilstudie ermittelten Feedback-Interaktionen auf. Diese weisen mit 42 % einen knapp doppelt so großen zeitlichen Anteil auf. Offenbar maßen die Lehrpersonen der TUFA-Stichprobe der Lernunterstützung in Feedback-Interaktionen deutlich mehr Gewicht bei als die Lehrpersonen der TIMS-Studie, die von Krammer untersucht worden waren. Nicht auszuschließen ist allerdings, dass die eher formale Operationalisierung einer Feedback-Interaktion in der Teilstudie von TUFA zu den höheren Werten geführt haben könnte.

6.2 Forschungsfrage 2: Qualität

Als Kernmerkmale der Qualität von formativem Assessment gelten die kognitive Aktivierung, der aktive Einbezug der Schülerinnen und Schüler sowie die ineinandergreifende Verbindung von diagnostischen Informationen und adaptiven Lernangeboten (Black und Wiliam 2009; Heritage 2007; Ruiz-Primo und Furtak 2006). Für jede der fünf Strategien wurden im Ratingmanual die Qualitätsansprüche entlang dieser Kernmerkmale für Qualität festgehalten. Jede Strategie wurde über Items mit vier- bzw. achtstufiger Skala geratet und anschliessend wurde pro Strategie ein Mittelwert gebildet. Insgesamt deuten die Ergebnisse des Ratings auf eine *geringe bis mittlere Qualität* hin. Die Beurteilungen anhand einer vierstufigen Skala ergaben erwartungswidrig tiefe Qualitätswerte. Die Qualität von *Peer-Assessment* wurde im Mittel am höchsten eingeschätzt ($M = 1{,}70$). Die Qualitätsratings für die Strategien *Lernziel* ($M = 1{,}05$) und *Self-Assessment* ($M = 1{,}14$) fielen noch tiefer aus. Die durchgeführten Faktorenanalysen zeigten, dass die Strategien *Elicting-Evidence* und *Feedback-Interaktion* zusammenfallen und dass zwischen dem Einsatz der Strategie *Eliciting-Evidence und Feedback-Interaktion in der Schülerarbeitsphase* und dem Einsatz der Strategie *Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht* unterschieden werden muss. Aus der Einschätzung dieser beiden Strategien auf einer achtstufigen Beurteilungsskala ($M = 3{,}19$ bzw. $M = 3{,}45$) wird ersichtlich, dass die Qualität der Strategieanwendung in beiden Fällen knapp unterhalb der Mitte liegt.

Diese Ergebnisse decken sich weitgehend mit den Befunden von Gotwals et al. (2015), die in ihrer Videostudie ebenfalls eine moderat ausgeprägte Qualität bei der Anwendung von Strategien des formativen Assessments feststellten. In dieselbe Richtung weisen die Ergebnisse der Beobachtungsstudie von Oswalt (2013), in der für die Strategien *Lernziel, Peer-Assessment* und *Self-Assessments* ebenfalls Qualitätsausprägungen im unteren und für die Strategien *Eliciting-Evidence* und *Feedback-Interaktion* Qualitätsausprägungen im mittleren Skalenbereich ermittelt wurden. Studien mit selbst berichteten Daten (z. B. Altmann et al. 2010; Schmidt 2020) kommen im Gegensatz dazu jedoch zu deutlich positiveren Ergebnissen, was als Hinweis darauf interpretiert werden kann, dass die Aussagekraft von Selbsteinschätzungen der Qualität der Umsetzung von formativen Assessments infrage zu stellen ist (Altmann et al. 2010; Schmidt 2020). Unabhängige externe Beobachterinnen und Beobachter, wie sie in TUFA und in den Studien von Gotwals et al. (2015) und Oswalt (2013) zum Einsatz kamen, gelangen offensichtlich zu kritischeren Urteilen, was die Qualität des formativen Assessments anbelangt, als die Lehrpersonen selbst. Diese Unterschiede in den Qualitätseinschätzungen dürften sich u. a. wesentlich auf unterschiedliche Erhebungsformate sowie auf Milde-Effekte in Studien mit Selbsteinschätzungen zurückführen lassen (Döring und Bortz 2016).

Bei der faktorenanalytischen Überprüfung des Ratinginstruments wurde festgestellt, dass die Strategien *Eliciting-Evidence* und *Feedback-Interaktion* – entgegen bestehenden theoretischen Annahmen (z. B. Black und Wiliam 2009) – keine eigenständigen Konstrukte darstellen. Durch diesen Befund wird die Annahme von Hattie und Timperley (2007) und Heritage (2010) gestützt, wonach bereits die mündliche Ermittlung des Lernstands (Eliciting-Evidence) als Feedback aufzufassen ist. Nach

Hattie und Timperley (2007) sind Feedbacks nicht auf bloße „Rückmeldungen“ zu reduzieren, sondern umfassen auch Fragen oder Impulse, die das Beschreiben von Lernprozessen und das Nachdenken darüber anregen.

Den dargestellten Ergebnissen zufolge bestehen zwischen den Qualitätsausprägungen der fünf Strategien teilweise Zusammenhänge. Es korrelieren die Skalen der Strategien *Eliciting-Evidence und Feedback-Interaktion in der Schülerarbeitsphase* und *Eliciting-Evidence und Feedback-Interaktion im Klassenunterricht* untereinander und diese beiden Strategien mit der Strategie *Peer-Assessment*. *Peer-Assessment* und *Self-Assessment* korrelieren auch miteinander. Das heißt, teilweise folgt aus einer hohen Qualität der Umsetzung einer Strategie auch eine hohe Qualität der Umsetzung einer anderen Strategie, jedoch nicht zwingend. Eine latente Profilanalyse könnte im Anschluss an die Forschungsergebnisse von Smit und Engeli (2017) darüber Auskunft geben, welche Strategien oder Kombinationen von Strategien die Lehrpersonen anwenden. Mit Bennett (2011) ist zudem zu fragen, ob das formative Assessment tatsächlich als ein einheitliches Konstrukt – gemessen anhand von fünf Strategien – verstanden werden kann. Im Anschluss an Bennett (2011) und Lyon et al. (2019) ist in künftigen Forschungsprojekten deshalb zu untersuchen, in welchem Verhältnis die fünf Strategien von Black und Wiliam (2009) zu sehen sind. Zu klären gilt es diesbezüglich beispielsweise, ob die verschiedenen Strategien zueinander in Konkurrenz stehen oder ob eine Strategie die andere kompensieren kann. Zudem stellt sich die grundsätzliche Frage, ob es tatsächlich genau diese fünf Strategien sind, die formatives Assessment auszeichnen, oder ob noch andere in Erwägung zu ziehen sind.

6.3 Forschungsfrage 3: Zusammenhang von Dauer und Qualität

Im letzten Analyseschritt wurde geklärt, inwiefern die Dauer und die Qualitätsausprägung einer Strategie zusammenhängen. Signifikante Zusammenhänge zwischen Dauer und Qualitätsausprägung finden sich bei allen Strategien. Die signifikant positiven Korrelationen sind u. a. im Hinblick auf die Unterrichtsplanung aufschlussreich, da sie darauf hinweisen, dass auch für qualitätsvolles formatives Assessment explizit Zeit eingeplant werden muss. Allerdings muss diese Schlussfolgerung vorsichtig zur Kenntnis genommen werden, denn das Ergebnis könnte auch auf die jeweiligen Operationalisierungen im Kodierinstrument und im Ratinginstrument zurückzuführen sein. Teilweise ist die zeitliche Quantität (gemäß Kodierung) eng mit der Qualitätseinschätzung (Rating) verbunden. Zur Erfassung der zeitlichen Quantität musste eine breite Operationalisierung gewählt werden, damit sämtliche Varianten (z. B. des Self-Assessments) erfasst werden konnten. Daraus resultierte jedoch eine konzeptionelle Ähnlichkeit der beiden Instrumente, was die Zusammenhänge zwischen Dauer und Qualität zumindest teilweise erklären könnte.

6.4 Schlussfolgerungen für die Aus- und Weiterbildung von Lehrpersonen

Die präsentierten Ergebnisse zur meist kurzen Einsatzdauer der Strategien und ihrer höchstens mittleren Qualität fielen eher erwartungswidrig aus, denn 70 % der an der Studie teilnehmenden Lehrpersonen waren zum Zeitpunkt der Videoaufnah-

men jünger als 36 Jahre und hatten weniger als zehn Jahre Berufserfahrung. Es ist daher zu vermuten, dass es sich bei der großen Mehrheit von ihnen um Absolventinnen und Absolventen von Pädagogischen Hochschulen gehandelt haben dürfte, die in der Schweiz explizit mit dem Ziel gegründet worden waren, die Professionalität der Lehrpersonen zu erhöhen. Daher hätte erwartet werden können, dass die Absolventinnen und Absolventen qualitätsvolles formatives Assessment realisieren. Offensichtlich gelingt es noch nicht in ausreichendem Maße, Lehrpersonen so aus- und weiterzubilden, dass sie in der Lage sind, qualitativ hochstehendes, lernförderliches formatives Assessment im Unterricht umzusetzen. Die Ergebnisse weisen zudem darauf hin, dass die untersuchten Lehrpersonen das Assessment noch nicht als essenziellen Bestandteil von qualitativ gutem Unterricht auffassen und auch noch nicht erkennen, dass sie insbesondere mithilfe von Self- und Peer-Assessments zentrale Aspekte von eigenständigem Lernen fördern könnten. Denkbar ist in diesem Zusammenhang jedoch auch, dass (jüngere) Lehrpersonen durch die bestehende Beurteilungskultur und Beurteilungspraxis in den Schulen, in denen sie unterrichten, in einer Weise beeinflusst werden, die es ihnen erschwert, das Lernerfolg erwarten lassende, jedoch anspruchsvolle Konzept von formativem Assessment, das ihnen an der Hochschule vermittelt wurde, im informellen alltäglichen Unterricht qualitativ zufriedenstellend umzusetzen (Bennett 2011). Zu prüfen ist daher, wie die Schulen vor Ort ihre Beurteilungskultur und ihre Beurteilungspraxis insgesamt optimieren könnten.

6.5 Einschränkungen, Ausblick und Fazit

Die berichteten Ergebnisse gelten grundsätzlich spezifisch für den Mathematikunterricht der vierten Primarklasse in der (Zentral-)Schweiz. Eine direkte Übertragung auf andere Fächer und Jahrgangstufen ist nur sehr beschränkt möglich, weil Hinweise darauf vorliegen, dass formatives Assessment domänenspezifisch ausfällt (Maier 2011) und in seiner Form auch zwischen den Jahrgangsstufen variieren kann (Bürgermeister 2014). Einschränkend muss außerdem darauf hingewiesen werden, dass die TUFA-Stichprobe im Vergleich mit anderen Videostudien (Gotwals et al. 2015; Lyon et al. 2019; Oswalt 2013; Ruiz-Primo und Furtak 2006) zwar als relativ groß angesehen werden kann, ihr Umfang für vertiefende Analysen bestimmter sich abzeichnender Muster von formativem Assessment jedoch trotzdem zu limitiert ist. Auch die Ergebnisse der Faktorenanalysen bedürfen der Replikation mit größeren Stichproben. Zudem ist bei der Zusammenstellung der Stichprobe von einer positiven Selektion auszugehen. Da die teilnehmenden Lehrpersonen damit einverstanden sein mussten, beim Unterrichten videografiert zu werden, war es kaum möglich, die mit Videostudien einhergehenden Limitationen hinsichtlich der positiven Auswahl zu umgehen.

Eine Überprüfung der Erkenntnisse zur Praxis des formativen Assessments in anderen Klassenstufen, in anderen Fächern und mit größeren Stichproben ist somit wünschenswert. Damit Ansatzpunkte für Veränderungen der gegenwärtigen Unterrichtspraxis gefunden werden können („Wie kommen Befunde der Wissenschaft in die Klassenzimmer?“, Lipowsky 2019), müsste geklärt werden, auf welchen Voraussetzungen ein qualitätsvolles und lernwirksames formatives Assessment basiert.

Diese Analyse wird im Rahmen der TUFA-Studie möglich sein, denn im Anschluss an die Videoaufnahmen füllten die teilnehmenden Schülerinnen und Schüler zusammen mit der Versuchsleiterin einen Papierfragebogen aus, der u. a. Items zur Fremdeinschätzung des formativen Assessments im Unterricht enthielt, während die Lehrpersonen in einem separaten Raum einen Fragebogen u. a. zum Professionswissen und zu berufsbezogenen Überzeugungen sowie zur Selbsteinschätzung des formativen Assessments bearbeiteten. Vor und nach den Videoaufnahmen hatten die Schülerinnen und Schüler zudem einen Leistungstest zum halbschriftlichen Dividieren gelöst, so dass noch weitere zusätzliche Daten miteinbezogen werden können. Auf der Grundlage von Leistungsdaten kann in Erweiterung zu den vorliegenden deskriptiven Befunden geprüft werden, ob das Ziel der Lernförderung (Black und Wiliam 2009) auch tatsächlich erreicht werden kann. Mit den Leistungsdaten können Aussagen dazu gemacht werden, welche Strategien oder Kombinationen von Strategien des formativen Assessments besonders lernförderlich für die Schülerinnen und Schüler sind. Zudem kann geprüft, ob alle oder nur bestimmte Gruppen von Lernenden (z. B. Schülerinnen und Schüler mit schwachen versus starken Schulleistungen) von formativem Assessment profitieren.

Zusammenfassend kann festgehalten werden, dass die vorliegenden Forschungsbefunde Hinweise darauf liefern, dass informelles formatives Assessment von einem großen Anteil der gefilmten Lehrpersonen im alltäglichen Unterricht durchgeführt wird. Die Anwendungsdauer der Strategien ist mit Ausnahme der Strategie *Feedback-Interaktionen* im Durchschnitt allerdings eher kurz bemessen. Zudem deuten die Ergebnisse des Ratings eher auf eine höchstens mittlere, meist aber geringe Qualität hin. Das Fehlen von hohen Qualitätsausprägungen legt die Schlussfolgerung nahe, dass es Lehrpersonen noch schwerfällt, metakognitive Prozesse anzuregen, Schülerinnen und Schüler durch offene, anregende Fragen kognitiv-aktivierend in Unterrichtsgespräche einzubeziehen, Lernprozesse adaptiv durch inhaltlich-fundierte Feedback-Interaktionen gezielt situativ voranzutreiben und weiterführende Denkprozesse anzuregen. Das Niveau scheint in der Regel eher einer oberflächlichen Beurteilung der dichotomen Form „gut/schlecht“, „richtig/falsch“ oder „kann ich/kann ich nicht“ zu entsprechen und weniger mit einem Beurteilungsverständnis einherzugehen, das darauf abzielt, eine fundierte Auseinandersetzung mit im Unterricht stattfindenden Lern- und Denkprozessen zu fördern.

Das Ziel muss sein, dass Lehrpersonen in der Lage sind, qualitätsvolle formative Assessments durchzuführen, beispielsweise indem sie prozessorientiert Lernziele und Erfolgskriterien festlegen und diese auch kommunizieren, Lernprozesse mit Fragen und lernwirksamem Feedback unterstützen und diese unter Rückbezug auf die zuvor explizit genannten Lernziele und Erfolgskriterien überprüfen (Harlen 2007). Eine solche qualitative Veränderung der gegenwärtig zu beobachtenden Praxis des formativen Assessments zugunsten des Lernerfolgs der Schülerinnen und Schüler kann jedoch nicht unabhängig vom übergeordneten Kontext der Unterrichtsentwicklung erfolgen, da das formative Assessment als zentraler Teil von Unterricht seine Wirkung nur im Verbund mit qualitätsvollem Unterricht entfalten kann.

Funding Open access funding provided by University of Teacher Education Lucerne

Anhang A

Beispiel für achtstufige Ratingskala

Beispiel Eliciting Evidence-Item EE-1 Schülerbeteiligung Die Schülerinnen und Schüler haben Zeit und Raum, um ihr Denken im Unterricht angemessen und ausführlich zu verbalisieren.

Hohe Ausprägung (6, 7):

1. Eigene Ideen: Viele (3 oder mehr) der Schülerinnen und Schüler haben Zeit und Raum, eigene Gedanken und Überlegungen zu entwickeln, diese einzubringen und auszuführen (längere Beiträge). Die Schülerinnen und Schüler haben einen hohen Sprechanteil (40–50 %).
2. Wartezeit: Die Lehrperson lässt den Schülerinnen und Schülern Zeit, um über Fragen nachzudenken (mehr als 2 Sek.). Klassenunterricht: auch, wenn sich jemand schon per Handzeichen meldet.
3. Unterbrechungen: Die Schülerinnen und Schüler können ihren Gedankengang ohne Unterbrechung ganz zu Ende führen.

Mittlere Ausprägung (3, 4, 5):

1. Eigene Ideen: Einige Schülerinnen und Schüler (2–3) haben die Gelegenheit, eigene Gedanken und Überlegungen einzubringen und diese auszuführen (längere Beiträge). Die Lehrperson hat den größten Sprechanteil, aber einige Schülerinnen und Schüler haben längere Beiträge.
2. Wartezeit: Die Lehrperson lässt den Schülerinnen und Schülern kurz Zeit (ca. 2 Sek.), um über die Fragen nachzudenken. Klassenunterricht: Bis der/die Erste sich per Handzeichen meldet, um zu antworten.
3. Unterbrechungen: Die Lehrperson wartet Antworten von Schülerinnen und Schülern in der Regel ab, ohne diese zu unterbrechen oder Antworten vorwegzunehmen.

Geringe Ausprägung (1, 2):

1. Eigene Ideen: Die Schülerinnen und Schüler haben keine Gelegenheit, eigene Gedanken und Überlegungen einzubringen. Die Schülerinnen und Schüler haben nur sehr geringe Gesprächsbeiträge (oftmals nur ein Wort). Die Lehrperson spricht fast ausschließlich selbst.
2. Wartezeit: Die Lehrperson lässt den Schülerinnen und Schülern zu wenig Zeit (<2 Sek.), um über die Fragen nachzudenken und/oder unterbricht Schülerinnen und Schüler.
3. Unterbrechungen: Die Lehrperson unterbricht die Schülerinnen und Schüler zum Teil oder nimmt Antworten vorweg bzw. beantwortet Fragen selbst.

Nicht einschätzbar (0)

Anhang B

Beispiel für vierstufige Ratingskala

Beispiel Lernziel-Item LZ-1 Vorwissen Die Lehrperson stellt beim Vorstellen der Lernziele Bezüge zum Vorwissen der Schülerinnen und Schüler her.

Hohe Ausprägung (3):

- Mehrere Aspekte des Vorwissens oder der Vorläuferfertigkeiten werden genannt und es wird deutlich, wie die neuen Inhalte mit bereits früher gelernten Inhalten zusammenhängen. Die Lehrperson lässt Schülerinnen und Schüler Bezüge zum Vorwissen herstellen.

Mittlere Ausprägung (2):

- Mehrere Aspekte des Vorwissens oder der Vorläuferfertigkeiten werden genannt, ohne dass dabei ein Zusammenhang mit den Lernzielen hergestellt wird.
- Oder: Ein Aspekt des Vorwissens wird genannt und für diesen wird ein Zusammenhang mit den Lernzielen verdeutlicht.

Geringe Ausprägung (1):

- Die Lehrperson geht höchstens auf einen Aspekt des Vorwissens kurz ein (Vorläuferfähigkeit oder vorangegangenen Unterricht). Die Schülerinnen und Schüler haben kaum oder keine Zeit, sich zu äußern. Zusammenhänge werden nicht deutlich.

Nicht einschätzbar (0)

Literatur

Altmann, P. C., Fleming, P. B., & Heyburn, S. L. (2010). Understanding and using formative assessments: A mixed methods study of assessment for learning adoption. Nashville: Vanderbilt University, Peabody College. https://pdfs.semanticscholar.org/ed5b/25023a3285a668fc48b6fef5d3596f0af900.pdf. Zugegriffen: 29. Febr. 2020.

Andrade, H. L. (2010). Students as the definitive source of formative assessment. Academic self-assessment and the self-regulation of learning. In H. L. Andrade & G. J. Cizek (Hrsg.), *Handbook of formative assessment* (S. 90–105). New York: Routledge. https://doi.org/10.4324/9780203874851.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education*. Dordrecht: Kluwer. https://doi.org/10.1002/sce.1022.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678.

Berner, N. E., Corvacho del Toro, I., Gabriel, K., & Denn, A.-K. (2013). Aufbereitung der Videodaten und Transkription. In F. Lipowsky & G. Faust (Hrsg.), *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE)* (S. 83–103). Frankfurt a.M.: GFPF.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–13. https://doi.org/10.1007/s11092-008-9068-5.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807.

Brookhart, S. M., Moss, C. M., & Long, B. A. (2010). Teacher inquiry into formative assessment practices in remedial reading classroom. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 41–58. https://doi.org/10.1080/09695940903565545.

Bürgermeister, A. (2014). *Leistungsbeurteilung im Mathematikunterricht. Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit*. Münster: Waxmann. https://doi.org/10.1080/09695940903565545.

Bürgermeister, A., & Saalbach, H. (2018). Formatives Assessment: Ein Ansatz zur Förderung individueller Lernprozesse. *Psychologie in Erziehung und Unterricht*, *65*(3), 194–205.

Cheng, L., & Wang, X. (2007). Grading, feedback and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, *4*(1), 85–107. https://doi.org/10.1080/15434300701348409.

Cizek, G. J. (2010). An introduction to formative assessment: history, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Hrsg.), *Handbook of formative assessment* (S. 3–17). New York: Routledge. https://doi.org/10.4324/9781315166933-1.

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

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting students' science understanding? *American Educational Research Journal*, *52*(6), 1133–1159. https://doi.org/10.3102/0002831215596412.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg, New York: Springer.

Gotwals, A. W., & Birmingham, D. (2016). Eliciting, identifying, interpreting, and responding to students' ideas: teacher candidates' growth in formative assessment practices. *Research in Science Education*, *46*(3), 365–388. https://doi.org/10.1007/s11165-015-9461-2.

Gotwals, A. W., Philhower, J., Cisterna, D., & Bennett, S. (2015). Using video to examine formative assessment practices as measures of expertise for mathematics and science teachers. *International Journal of Science and Mathematics Education*, *13*(2), 405–423. https://doi.org/10.1007/s10763-015-9623-8.

Groeben, N., Wahl, D., Schlee, J., & Scheele, B. (1988). *Forschungsprogramm Subjektive Theorien. Eine Einführung in die Psychologie des reflexiven Subjekts*. Tübingen: Francke.

Harlen, W. (2007). Formative classroom assessment in science and mathematics. In J. H. McMillan (Hrsg.), *Formative classroom assessment. theory into practice* (S. 116–135). New York: Teachers College Press. https://doi.org/10.1007/s11618-010-0124-9.

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: case studies into teachers' implementation. *Teaching and Teacher Education*, *36*, 101–111. https://doi.org/10.1016/j.tate.2013.07.008.

Hattie, J. (2003). *Formative and summative interpretations of assessment information*. Auckland: University of Auckland, School of Education.

Hattie, J. (2016). *Lernen sichtbar machen*. Baltmannsweiler: Schneider Hohengehren. Überarbeitete deutschsprachige Ausgabe von „Visible Learning" besorgt von Wolfgang Beywl und Klaus Zierer

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487.

Heritage, M. (2007). Formative assessment: what do teachers need to know and do? *Phi Delta Kappan*, *89*(2), 140–145. https://doi.org/10.1177/003172170708900210.

Heritage, M. (2010). *Formative assessment: making it happen in the classroom*. Thousand Oaks: Corwin Press.

Hugener, I. (2006). Kapitel 3: Überblick über die Beobachtungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Teil 3: Videoanalysen* (S. 45–54). Frankfurt a.M.: DIPF.

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung – eine Videoanalyse*. Münster: Waxmann.

Klauer, K. J. (2014). Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik. In M. Hasselhorn, W. Schneider & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 1–17). Göttingen: Hogrefe.

Klieme, E., Pauli, C., & Reusser, K. (2006). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Teil 3: Videoanalysen*. Frankfurt a.M.: DIPF.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254.

Kobarg, M., & Seidel, T. (2003). Prozessorientierte Lernbegleitung im Physikunterricht. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"* (S. 151–200). Kiel: IPN.

Krammer, K. (2009). *Individuelle Lernunterstützung in Schülerarbeitsphasen. Eine videobasierte Analyse des Unterstützungsverhaltens von Lehrpersonen im Mathematikunterricht*. Münster: Waxmann.

Krammer, K., & Hugener, I. (2014). Förderung der Analysekompetenz angehender Lehrpersonen anhand von eigenen und fremden Unterrichtsvideos. *Journal für Lehrerinnen- und Lehrerbildung*, *14*(1), 25–32.

Lipowsky, F. (2019). Wie kommen Befunde der Wissenschaft in die Klassenzimmer? – Impulse der Fortbildungsforschung. In C. Doni, F. Foerster, M. Obermayr, A. Deckwerth, G. Kammermeyer, G. Lenske, M. Leuchter & A. Wildemann (Hrsg.), *Grundschulpädagogik zwischen Wissenschaft und Transfer* (S. 144–161). Wiesbaden: Springer VS.

Lotz, M. (2016). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr*. Wiesbaden: Springer VS.

Lotz, M., Berner, N. E., Gabriel, K., Post, S., Faust, G., & Lipowsky, F. (2011). Unterrichtsbeobachtung im Projekt PERLE. In D. Kucharz, T. Irion & B. Reinhoffer (Hrsg.), *Grundlegende Bildung ohne Brüche* (S. 183–194). Wiesbaden: VS. https://doi.org/10.1007/978-3-531-94131-8_34.

Lyon, J., Nabors Olah, L., & Wylie, C. (2019). Working toward integrated practice: Understanding the interaction among formative assessment strategies. *The Journal of Educational Research*, *112*(3), 301–314. https://doi.org/10.1080/00220671.2018.1514359.

Maier, U. (2011). Formative Leistungsdiagnostik in der Sekundarstufe I – Befunde einer quantitativen Lehrerbefragung zu Nutzung und Korrelaten verschiedener Typen formativer Diagnosemethoden in Gymnasien. *Empirische Pädagogik*, *25*(1), 25–46.

Oswalt, S. G. (2013). Identifying formative assessment in classroom instruction: creating an instrument to observe use of formative assessment in practice. Dissertation. Boise: Boise State University. https://scholarworks.boisestate.edu/cgi/viewcontent.cgi?article=1772&context=td. Zugegriffen: 10. Okt. 2018.

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The future of student self-assessment: a review of known unknown and potential directions. *Educational Psychology Review*, *28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2.

Pauli, C. (2012). Kodierende Beobachtung. In H. Boer & S. Reh (Hrsg.), *Beobachtung in der Schule – Beobachten lernen* (S. 45–63). Wiesbaden: VS. https://doi.org/10.1007/978-3-531-18938-3_3.

Pellegrino, J. W., Chudosky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. Washington: National Academy Press.

Petko, D. (2006). Kameraskript. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Teil 3: Videoanalysen* (S. 15–37). Frankfurt a.M.: DIPF.

Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, *31*(121), 33–46. https://doi.org/10.1002/yd.295.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom assessment scoring system. Upper elementary manual*. Charlottesville: Teachstone.

Reusser, K. (1995). Lehr-Lernkultur im Wandel: Zur Neuorientierung in der kognitiven Lernforschung. In R. Dubs & R. Dörig (Hrsg.), *Dialog & Wissenschaft & Praxis* (S. 164–190). St. Gallen: IWP.

Rimmele, R. (2004). *Videograph. Multimedia-Player zur Kodierung von Videos*. Kiel: IPN.

Ruiz-Primo, M. A. (2011). Informal formative assessment: the role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, *37*(1), 15–24. https://doi.org/10.1016/j.stueduc.2011.04.003.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: exploring teachers' practices and student learning. *Educational Assessment*, *11*(3–4), 205–235. https://doi.org/10.1080/10627197.2006.9652991.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714.

Schmidt, C. (2020). *Formatives Assessment in der Grundschule*. Wiesbaden: Springer.

Schmidt, C., & Liebers, K. (2015). Formatives Assessment an Grundschulen – Praxis und Bedingungsfaktoren. In K. Liebers, B. Landwehr, A. Marquardt & K. Schlotter (Hrsg.), *Lernprozessbegleitung und adaptives Lernen in der Grundschule. Forschungsbezogene Beiträge* (S. 133–138). Wiesbaden: Springer VS.

Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. Forum: Qualitative Sozialforschung, 15(1), Artikel 18. http://www.qualitative-research.net/index.php/fqs/rt/printerFriendly/2043/3635. Zugegriffen: 29. Febr. 2020.

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*(4), 697–715. https://doi.org/10.1007/s11858-017-0855-7.

Seidel, T., Prenzel, M., Duit, R., & Lehrke, M. (Hrsg.). (2003). *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"*. Kiel: IPN.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. *Applied Measurement in Education*, *21*(4), 295–314. https://doi.org/10.1080/08957340802347647.

Smit, R. (2008). Formative Beurteilung im kompetenz- und standardorientierten Unterricht. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *26*(3), 383–392.

Smit, R., & Engeli, E. (2017). Formative Beurteilung im jahrgangsübergreifenden Unterricht. *Zeitschrift für Erziehungswissenschaft*, *20*(2), 279–303. https://doi.org/10.1007/s11618-016-0697-z.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap*. New York: Free Press.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303.

Topping, K. J. (2010). Peers as a source of formative assessment. In H. L. Andrade & G. J. Cizek (Hrsg.), *Handbook of formative assessment* (S. 61–74). New York: Routledge. https://doi.org/10.1080/713611428.

Trepke, C., Seidel, T., & Dalehefte, I. M. (2003). Zielorientierung im Physikunterricht. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht"* (S. 201–229). Kiel: IPN.

Widmer-Wolf, P., Sieber-Suter, B., & Thierstein, C. (Hrsg.). (2014). Eine Sammlung berufsspezifischer Kompetenzen für das Berufsfeld Schule. Bruu-Windisch: Pädagogische Hochschule FHNW, Institut Weiterbildung und Beratung. https://scenario.kompetenzmanager.ch/myUploadData/files/sammlung_berufsspezifische_kompetenzen_14_v4.pdf. Zugegriffen: 23. März 2020.

Wylie, C., & Lyon, C. (2013). Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice. Washington: FAST, SCASS & CCSSO. https://www.pblconnections.com/uploads/2/6/7/2/26722395/formative_assessment_rubrics_and_observation_tools_document.pdf. Zugegriffen: 23. März 2020.