Empirische Sonderpädagogik, 2022, Nr. 4, S. 368-404
ISSN 1869-4845 (Print) · ISSN 1869-4934 (ebook)

Leistungsbewertung in inklusiven Lernkontexten: Wie beurteilen Grundschullehrkräfte die Eignung alternativer Formen der Leistungsbewertung in heterogenen Lerngruppen?

Henrike Kopmann, Horst Zeinz und Magdalena Kaul

Westfälische Wilhelms-Universität Münster

Zusammenfassung
Durch die sukzessive Umsetzung schulischer Inklusion stellt sich zunehmend die Frage nach geeigneten Maßnahmen, welche die individuelle Leistung von Kindern mit und ohne besonderen Förderbedarf diagnostizieren und unterstützen können. Dabei kommen vermehrt alternative Formen der Leistungsbeurteilung in den Blick. In der vorliegenden Studie wird untersucht, inwiefern diese bei Grundschullehrkräften bekannt und gerade im Hinblick auf den Einsatz für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf akzeptiert sind. Im Rahmen einer Querschnittsstudie wurden 40 Lehrkräfte an 13 Grundschulen mittels Fragebögen untersucht. Insgesamt waren die genannten alternativen Formen der Leistungsbewertung bekannt und wurden auch als geeignet für die Leistungsbeurteilung von Schülerinnen und Schülern eingeschätzt. Als besonders geeignet für Lernende mit sonderpädagogischem Förderbedarf wurden Lernportfolios angesehen. Kompetenzraster und Schülerinnen- und Schülerselbstbewertungen wurden ebenso positiv eingeschätzt, wohingegen Gruppenarbeitsergebnisse insgesamt als eher ungeeignet angesehen wurden. Korrelationsanalysen zeigten, dass Lehrkräfte, die Lernportfolios, Schülerinnen- und Schülerselbstbewertungen und Gruppenarbeitsergebnisse als geeignete Formen der Leistungsbeurteilung von Lernenden mit sonderpädagogischem Förderbedarf ansahen, auch statistisch signifikant häufiger von einem Einsatz der jeweiligen Form im Unterricht berichteten. Limitationen der Studie sowie mögliche Implikationen der Befunde für weitere Forschungsprojekte, für die Lehrkräftebildung und für die Schulpraxis werden diskutiert.

Schlüsselwörter: Leistungsbewertung, Inklusion, Grundschule, Lehrkräfte

Performance assessment in inclusive learning contexts: How do primary school teachers evaluate alternative forms of performance assessment in heterogeneous student groups?

Abstract
The successive implementation of school inclusion increasingly raises the question of suitable measures that can diagnose and support the individual performance of children with

and without special needs. Alternative forms of performance assessment are increasingly being considered. The present study examines to what extent these are known to primary school teachers and to what extent they are accepted, especially with regard to pupils with special needs. As part of a cross-sectional study, 40 teachers at 13 primary schools were examined using questionnaires. Overall, the alternative forms of assessment mentioned were known and were also considered as suitable for assessing the performance of pupils. Learning portfolios were seen as particularly suitable for pupils with special educational needs. Competence frames and student self-assessments were also rated positively, whereas group work results were generally viewed as rather unsuitable. Correlation analyses showed that teachers who regarded learning portfolios, student self-assessments and group work results as suitable forms of performance assessment of students with special educational needs also reported that the respective form was used more frequently in the classroom. Limitations of the study and possible implications of the findings for further research projects, for teacher training and for school practice are discussed.

Keywords: performance assessment, inclusion, primary school, teachers

Die gesellschaftlich und bildungspolitisch propagierte Leitidee der Inklusion fordert einen veränderten, von Wertschätzung und Chancengerechtigkeit geprägten Umgang mit menschlicher Diversität (Amrhein & Reich, 2014). Im schulischen Kontext werden diesbezüglich eine förderdiagnostisch fundierte Individualisierung und Adaption von Lernangeboten an die jeweiligen Lernvoraussetzungen der Schülerinnen und Schüler gefordert (u.a. Buholzer et. al., 2014; Textor, 2015). Die traditionelle schulische Leistungsbewertung über Ziffernnoten steht jedoch in einem Widerspruch zu einer heterogenitätssensiblen und motivierenden individuellen Förderung (Streese et al., 2017). Dies gilt insbesondere für Schülerinnen und Schüler, deren Leistungen – vor dem Hintergrund sozialer oder kriterialer Bezugsnormen – als gering bewertet werden. Angesichts des aktuellen Inklusionsdiskurses stellt sich daher unweigerlich die Frage nach differenzbejahenden, alternativen Formen schulischer Leistungsbewertung (Bohl, 2019; von Bargen, 2017). Der nachfolgende Beitrag greift die entsprechende Problematik auf: Nach der Betrachtung unterschiedlicher Funktionen schulischer Leistungsbewertung werden mögliche Potenziale, aber auch Limitationen alternativer Formen der Leistungsbewertung in inklusiven Lernkontexten herausgearbeitet. Ausgehend vom empirischen Forschungsstand eruiert die hier vorgestellte Studie die Eignung alternativer Formen der Leistungsbewertung aus der Sicht von Grundschullehrkräften.

Funktionen schulischer Leistungsbewertung

Die schulische Leistungsbewertung erfüllt vielfältige und teilweise widersprüchliche Funktionen, wobei zwischen gesellschaftlichen und pädagogischen Zielsetzungen differenziert wird (Fürstenau & Gomolla, 2012; Lintorf, 2012; Wild & Krapp, 2006).

Gesellschaftliche Funktionen

In gesellschaftlicher Hinsicht sind die Allokations- und Selektionsfunktion hervorzuheben: Eine Quantifizierung von Leistungen, z.B. in Form der traditionellen Ziffernnote, kontrolliert den Zugang zur nächsten Klassenstufe, zu unterschiedlichen Schulformen sowie zu Ausbildungs-, Studien- und Arbeitsplätzen (Lintorf, 2012). Bewerberinnen und Bewerber werden ausgelesen, wobei schulische Leistungen

spätere Erfolge prognostizieren (Prognosefunktion) und Selektionsentscheidungen rechtfertigen sollen. Schulische Leistungen gelten als allgemein anerkanntes Kriterium für die Verteilung beruflicher Möglichkeiten, hiermit verbundener finanzieller Ressourcen und Statusaspekte (Bräu, 2018).

Das meritokratische Prinzip, wonach leistungsstarke Personen mit besseren Aufstiegschancen belohnt werden, bleibt hierbei zumeist unhinterfragt (Bräu, 2018). In diesem gesellschaftlichen Kontext hat die schulische Benotungspraxis auch die Funktion, Heranwachsende dahingehend zu sozialisieren, dass sie das Leistungsprinzip – im Sinne einer leistungsabhängigen Verteilung gesellschaftlicher Positionen und Chancen – als gerecht anerkennen und verinnerlichen (Lintorf, 2012; Wild & Krapp, 2006). Die Fairness des gesellschaftlichen Leistungsprinzips ist jedoch auch kritisch zu betrachten (Falkenberg, 2020; Fürstenau & Gomolla, 2012): So entscheiden nicht nur die individuelle Anstrengung und Fähigkeit über das abschließende Leistungsprodukt bzw. den Erfolg. Zufall, Selbstdarstellung und förderliche Beziehungen spielen ebenfalls eine entscheidende Rolle (Bräu, 2018). Weiterhin nehmen v.a. bildungssprachliche Kompetenzen sowie das finanzielle und kulturelle Kapital der Ursprungsfamilie Einfluss auf Schulnoten. Hierbei kann die schulische Beurteilungspraxis insbesondere die Bildungsungleichheit und Benachteiligung von Lernenden mit Migrationshintergrund oder geringem sozioökonomischen Status oder beidem zementieren (Allemann-Ghionda et al., 2006; Falkenberg, 2020). Zudem ist Leistung im pädagogischen und insbesondere inklusionspädagogischen Kontext eine Perspektivfrage (Fürstenau & Gomolla, 2012): Ist eher die individuelle Anstrengung und Einsatzbereitschaft oder aber das erreichte Leistungsergebnis zu bewerten? Sind bei der Bewertung einseitig kognitive Leistungsaspekte zu betonen oder weitere Kompetenzen einzubeziehen, z.B. im Bereich der sozialen Kooperation oder des selbstregulierten Arbeitens?

Eine weitere gesellschaftliche Funktion der schulischen Leistungsbewertung kann in der Rechenschaftsablegung des Bildungs- und Erziehungswesens gegenüber der Öffentlichkeit gesehen werden (Fürstenau & Gomolla, 2012). Die Leistungen von Lerngruppen dienen hierbei als Nachweis erfolgreichen schulischen Lernens und legitimieren die entsprechende Form der institutionalisierten Wissensvermittlung (Lintorf, 2012). Groß angelegte Vergleichsarbeiten auf nationaler und internationaler Ebene (z.B. VERA, IGLU, PISA, TIMMS) stellen in diesem Kontext besondere Formen der schulischen Leistungsbewertung dar, wobei v.a. die Leistungsfähigkeit von Schulen im nationalen und internationalen Vergleich im Fokus stehen (Lintorf, 2012; Maier, 2015). Diese im englischen Sprachraum als *large-scale-assessments* bezeichneten Erhebungen nutzen standardisierte, schulextern erstellte Kompetenztests, welche v.a. den psychometrischen Gütekriterien der Objektivität, Reliabilität und Validität gerecht werden sollen (Heydrich et al., 2013). Die Angemessenheit entsprechender Leistungsmessungen für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf wird in der Literatur wiederholt diskutiert (Pohl et al., 2016). Der Einbezug von Lernenden mit besonderem pädagogischen Unterstützungsbedarf birgt methodische Herausforderungen, was eine adäquate Adaption von Testmaterialien angeht. Diesbezüglich wird u.a. auf eine angepasste Bearbeitungszeit und Möglichkeiten der mündlichen oder schriftlichen Beantwortung verwiesen (Mitchell et al., 2010). Zusätzlich werden alternative Formen der Kompetenzerfassung für Lernende mit sonderpädagogischem Förderbedarf angedacht, wie Checklisten, Portfolios oder die Sammlung repräsentativer Arbeitsprodukte (Mitchell, 2015). Maßstab der Bewertung solle hierbei die Erreichung individualisierter Leistungskriterien sein, die zuvor in einem standardorientierten individuellen Förderplan zu spezifizieren sind (Mitchell, 2015). Angesichts der skizzierten indivi-

duellen Anpassungen ist die Vergleichbarkeit und Objektivität der entsprechenden Leistungserhebungen allerdings kritisch zu hinterfragen (Heydrich et al., 2013). Nutzen und Sinnhaftigkeit einer Teilnahme von Schülerinnen und Schülern mit sonderpädagogischem Unterstützungsbedarf an nationalen und internationalen Vergleichsarbeiten sind ferner unter leistungsbezogenen und sozio-emotionalen Entwicklungsaspekten zu reflektieren: So kann die regelhafte Teilnahme an Vergleichsarbeiten ein am Regelcurriculum orientiertes, anspruchsvolles Unterrichtsangebot und Leistungszuwächse der betreffenden Lernenden begünstigen (Mitchell et al., 2010). Allerdings verweisen Einzelstudien auch auf negative Effekte, wie z.B. eine Zunahme von Misserfolgserfahrungen und leistungsbezogenen Sorgen sowie ein geringes Selbstbewusstsein bei Lernenden mit sonderpädagogischem Unterstützungsbedarf, die z.B. in den USA verpflichtend an *large-scale-assessments* teilnehmen (Mitchell, 2015).

Pädagogische Funktionen

Die genuin pädagogischen Funktionen schulischer Leistungsbewertung beziehen sich vor allem auf eine Optimierung von Lehr- und Lernprozessen (Lintorf, 2012). Die pädagogische Rückmeldefunktion betrifft hierbei unterschiedliche Zielgruppen: Lernende sollen über die Rückmeldung ihrer Leistungen dazu befähigt werden, eigene Lernergebnisse realistisch einzuschätzen und Lernprozesse selbstreguliert zu steuern (Maier, 2015; Schmidinger, 2013). Gegenüber den Eltern erfüllt die Leistungsdokumentation eine Berichtsfunktion und hilft ihnen im Idealfall dabei, ihr Kind zu unterstützen. Zudem dienen schulische Bewertungen sowohl den Schülerinnen und Schülern als auch ihren Eltern als Orientierung bei schulischen und beruflichen Laufbahnentscheidungen (Wild & Krapp, 2006). Lehrkräften bietet die Dokumentation der Leistungsentwicklung in der Lerngruppe oder bei einzelnen Lernenden Anhaltspunkte, um den eigenen Unterricht zu reflektieren und zu optimieren (Lintorf, 2012).

Überdies wird der schulischen Leistungsbewertung eine Motivationsfunktion zugeschrieben. Sie stelle einen extrinsischen Anreiz für Schülerinnen und Schüler dar, sich mit Inhalten auseinanderzusetzen, die unter Umständen wenig Bezug zu ihren intrinsischen Interessen aufweisen. Negative Bewertungen wiederum fungierten als Mittel der Disziplinierung (Lintorf, 2012; Wild & Krapp, 2006).

Konkret bezogen auf Lernende mit Lernschwierigkeiten und Verhaltensauffälligkeiten sind ferner folgende pädagogische Funktionen der Leistungsmessung und -bewertung besonders hervorzuheben: (1.) die statusdiagnostische Feststellung des aktuellen Leistungsstandes, (2.) die Identifikation möglicher Ursachen eines Leistungsversagens, (3.) eine diagnostisch fundierte Ableitung von Interventionen, (4.) die Auswahl von Fördermaßnahmen oder geeigneter Bildungsgänge oder beidem und (5.) die Evaluation der eingesetzten Fördermaßnahmen (Lenhard, 2014).

Formatives und summatives Assessment

Im englischsprachigen Diskurs findet sich die Unterscheidung zwischen einem summativen und formativen Assessment (Maier, 2015; Schmidt, 2020). Eine summative Bewertung, im Englischen als *summative assessent* oder *assessment of learning* bezeichnet, bezieht sich auf eine abschließende Beurteilung eines zu einem bestimmten Zeitpunkt erreichten Leistungsstandes (Schmidinger et al., 2016; Stiggins et al., 2007). Es handelt sich folglich um eine statusdiagnostische Beurteilung des Leistungsstandes, die z.B. als Basis von Selektionsentscheidungen herangezogen werden kann. Eine formative Bewertung, im Englischen als *formative assessment* oder *assessment for learning* bezeichnet, dient der Optimierung laufender Lehr- und Lernprozesse (Stiggins et al., 2007). Im Sinne einer

Prozessdiagnostik werden Lernleistungen wiederholt gemessen und bewertet, um das Lernen der Schülerinnen und Schüler über gezielte unterrichtliche Adaptionen zu fördern (Maier, 2015).

In der Praxis ist eine eindeutige Differenzierung zwischen summativen und formativen Formen der Leistungsbewertung häufig nicht möglich und beide Bewertungsaspekte schließen sich nicht gegenseitig aus (Schmidt, 2020). Zudem unterscheiden sich formative und summative Bewertungsformen nicht zwangsläufig in den eingesetzten Methoden (Maier, 2015; Schmidinger et al., 2016). Beispielsweise beeinflusst ein summativ bewertendes Abschlusszeugnis nachfolgende Laufbahnentscheidungen, Ausbildungs-, Studien- und Berufsmöglichkeiten. Hierin spiegelt sich einerseits eine selektive Allokationsfunktion wider. Andererseits werden aber auch formative Aspekte tangiert, indem ein passendes Lern- und Berufsumfeld ausgewählt wird und Lernbedingungen längerfristig optimiert werden. Ebenso bieten Klassenarbeiten – als klassische Formen der Leistungsmessung – in der Regel nicht nur eine summative Bewertung in Form einer Ziffernnote, sondern die Kennzeichnung von Fehlern und inhaltliche Hinweise der korrigierenden Lehrperson können zur formativen Optimierung zukünftiger Lernprozesse genutzt werden, z.B. über eine genauere Fehleranalyse.

Zur differenzierteren Betrachtung des Verhältnisses zwischen dem summativen und formativen Assessment lohnt sich ein Blick in den sonderpädagogischen Diskurs um eine Status- und Prozessdiagnostik (Neumann & Lütje-Klose, 2020; Reichenbach & Tiemann, 2018): Eine initiale Statusdiagnostik, welche einen Lernenden im Vergleich zu Gleichaltrigen und in Bezug auf spezifische Leistungs- und Entwicklungskriterien einstuft und einen Leistungsstand summativ feststellt, bildet dabei den Ausgangspunkt einer fundierten Förderplanung (Gold, 2011). Die anschließende wiederholte Messung der Lernfortschritte von Lernenden unter den entsprechenden pädagogischen Fördermaßnahmen entspricht wiederum einer förderorientierten Prozessdiagnostik, welche – wie auch das formative Assessment – Diagnostik und schulische Förderung systematisch verknüpft (Schmidt, 2020). Nachdem pädagogische Interventionen über eine bestimmte Zeit realisiert wurden, sind wiederum eher summative Bewertungselemente denkbar, um den abschließenden Erfolg einer Fördermaßnahme zu evaluieren. Wie ersichtlich greifen formative und summative Aspekte der Leistungsbewertung in der Praxis ineinander.

Inklusion und schulische Leistungsbewertung

Inklusionsdidaktische Publikationen vertreten ein primär förderorientiertes Verständnis schulischer Leistungsbewertung, in dem Beurteilungsprozesse in ein formatives pädagogisches Gesamtkonzept eingebettet sind (Buholzer et al., 2014; Textor, 2015; Schmidt & Liebers, 2017). Charakteristische Spannungsfelder zwischen der Forderung nach einem inklusiven, individuell fördernden Unterricht und schulischen Rahmenbedingungen manifestieren sich insbesondere im Bereich der Leistungsbewertung (Dietrich, 2017; Joller-Graf, 2010): Einerseits richtet sich das Augenmerk der inklusiven Agenda auf die Kultivierung einer gleichberechtigten und nicht hierarchisierenden Vielfalt (Beutel & Pant, 2019). Andererseits befördern Leistungsvergleiche und leistungsbezogene Kategorisierungen im schulischen Kontext unweigerlich soziale Vergleiche und Konkurrenz.

Tabelle 1 vermittelt einen Überblick über Antinomien im Spannungsfeld inklusionsdidaktischer Forderungen und traditioneller Formen der schulischen Leistungsbewertung (Kopmann, 2016).

Tabelle 1
Exemplarische Spannungsfelder zwischen inklusionsdidaktischen Forderungen und traditionellen Formen schulischer Leistungsbewertung

Inklusionsdidaktische Forderungen	Traditionelle schulische Leistungsbewertung
Individuelle Förderung orientiert an persönlichen Lern- und Entwicklungsvoraussetzungen	Orientierung an Lehrplanvorgaben, zunehmende Standardisierung
Binnendifferenzierte Lernangebote	Streben nach Objektivität von Beurteilungen und interpersonaler Vergleichbarkeit
Betonung der individuellen Bezugsnorm	Betonung der sozialen Bezugsnorm
Formative Leistungsdiagnostik	Summative Leistungsbewertung
Weiter Lern- und Leistungsbegriff umfasst z.B. sozio-emotionale, musische und kognitive Dimensionen	Enger Lern- und Leistungsbegriff fokussiert auf kognitive Dimension
Ganzheitliches Lernen	Fokussierung auf messbare schulische Leistungen
Freie Persönlichkeitsentfaltung	Vorbereitung auf gesellschaftliche Anforderungen
Erfolgserlebnisse für alle Lernenden	Selektions- und Allokationsfunktion von Schule
Anerkennung und Wertschätzung in demokratischer und nicht hierarchischer Gemeinschaft	Kategorisierende Leistungsbewertung als Monopol der Lehrperson

Zunächst stehen inklusionsdidaktische Ansprüche an schulisches Lernen, wie z.B. eine individuelle Förderung unter Berücksichtigung unterschiedlicher Lernvoraussetzungen und die Binnendifferenzierung von Unterricht, im Widerspruch zu einheitlichen Lehrplanvorgaben und einer Ausrichtung auf standardisierte und interpersonal vergleichbare Leistungsmessungen, z.B. durch Vergleichsarbeiten (Ainscow et al., 2006). Die Vereinbarkeit von Inklusion und kompetenzorientierten Bildungsstandards im Sinne normierter Leistungserwartungen wird in der Literatur wiederholt aufgegriffen (Holder & Kessels, 2019). Die entsprechenden Spannungsfelder werden häufig polarisierend – als inkompatible Gegenpole – dargestellt. Der Kompetenzbegriff darf jedoch nicht verkürzt, als einseitig kognitive Leistungserwartung verstanden werden (Frohn, 2019). Seine theoretische Konzeptualisierung ist wesentlich facettenreicher und beinhaltet – neben der kognitiven Komponente – z.B. soziale und motivationale Aspekte sowie praktische und ethische Handlungsaspekte (Weinert, 2001). Unter inklusiver Perspektive erscheint insbesondere die Definition von prozessorientierten Standards einer inklusiven Unterrichtsführung erstrebenswert (Frohn, 2019; Schuck, 2014).

Im Kontext einer inklusiven Leistungsbewertung werden ferner unterschiedliche Bezugsnormen kritisch diskutiert. In inklu-

sionsdidaktischen Publikationen findet sich häufig das Ideal einer individuellen Bezugsnorm schulischer Leistungsbewertung (Textor, 2015; Walm et al., 2017). Hierbei werden frühere Leistungen der Lernenden mit späteren Leistungen intraindividuell verglichen, wodurch individuelle Lernfortschritte verdeutlicht werden sollen (Reichenbach & Thiemann, 2018). Neben dieser individuellen Bezugsnorm plädieren u.a. Fischbach et al. (2021) sowie Mitchell et al. (2010) für eine individuell angepasste, sachliche Bezugsnorm, bei der Bewertungskriterien an unterschiedliche Lernausgangslagen und individualisierte Zielspezifikationen geknüpft werden. In der Schulpraxis werden pädagogische Handlungsspielräume hinsichtlich einer individualisierten oder individuell angepassten sachlichen Leistungsbewertung jedoch durch rechtliche Vorgaben limitiert, z.B. zur zielgleichen und zieldifferenten Beschulung in unterschiedlichen Förderschwerpunkten (Streese et al., 2017).

Traditionelle Formen der schulischen Leistungsbewertung legen zumeist vereinheitlichte sachliche Bezugsnormen zugrunde, indem Lernziele an curriculare Vorgaben angelehnt werden (Schmidinger, 2012). In der Praxis wird die sachliche Bezugsnorm jedoch von der sozialen Bezugsnorm in der Klasse überlagert, bei der Bewertungen durch interpersonale Vergleiche zwischen Schülerinnen und Schülern beeinflusst werden (Reichenbach & Thiemann, 2018). Wie Studien zeigen, geben Noten daher primär Auskunft über leistungsabhängige Rangfolgen in einer Klasse, sind klassenübergreifend aber kaum vergleichbar (Jachmann, 2003; Südkamp et al., 2012).

Weiterhin fokussieren inklusionsorientierte Publikationen häufig eine formative Leistungsbewertung (Textor, 2015; Walm et al., 2017). In diesem Kontext kritisieren u.a. Walm et al. (2017), dass traditionelle Formen der schulischen Leistungsbewertung in zu geringem Maße rückgebunden seien an weiterführende Planungsprozesse im Unterricht. In ihrem Modell einer erweiterten und integrierten pädagogischen Diagnostik rekurrieren die Autorinnen und Autoren auf alternative Formen der Leistungsbewertung nach Winter (2012). Eine Transparenz hinsichtlich der Kriterien der Leistungsbeurteilung, die Kommunikation und gemeinsame Reflexion über Lernprozesse und -ergebnisse sowie die Partizipation der Lernenden an Beurteilungsprozessen spielen hierbei eine wesentliche Rolle (Walm et al., 2017). Die Förderung der Kompetenz zur Bewertung eigener Leistungen durch die Schülerinnen und Schüler wird als Lernprozess an sich, als *assessment as learning*, angesehen (Walm et al., 2017).

Weiterhin vertreten inklusionsdidaktische Modelle zumeist einen weiten und ganzheitlichen Lern- und Leistungsbegriff, der z.B. sozio-emotionale, musische und kognitive Dimensionen umfasst (Kahlert & Heimlich, 2012; Textor, 2015). Eine traditionelle schulische Leistungsbewertung fokussiert hingegen auf die kognitive Dimension schulischen Lernens sowie auf klar messbare schulische Leistungen (Arndt & Werning, 2017; Sturm, 2015). Weiterführend stehen hinter den beschriebenen Lern- und Leistungsbegriffen unterschiedliche Ziele schulischer Bildung: Eine freie Persönlichkeits- und Potenzialentfaltung sowie die motivierende Würdigung kindlicher Lernanstrengungen für alle Lernenden werden durch inklusive Leitideen besonders betont (Streese et al., 2017). Traditionelle Formen der schulischen Leistungsbewertung akzentuieren hingegen eher die Vorbereitung auf gesellschaftliche Anforderungen sowie die Selektions- und Allokationsfunktion von Schule (Schmidinger, 2013).

In einem gesamtgesellschaftlichen Kontext spiegeln sich in unterschiedlichen Formen der Leistungsbewertung – oder eben auch in einem Verzicht darauf – normative Grundfragen des sozialen Miteinanders wider (Beutel & Pant, 2019): Inklusive Leitideen betonen die Anerkennung und Wertschätzung in einer demokratischen und nicht hierarchischen Gemeinschaft (Prengel, 2006). Das Monopol der Lehrkraft in puncto Leistungsbewertung wird hinterfragt

(Seitz & Simon, 2014; Walm et al., 2017).

Klassische Formen der Leistungsbewertung sehen hingegen eine kategorisierende und intersubjektiv vergleichende Leistungsbewertung durch die Lehrperson vor. Diese intersubjektive Vergleichbarkeit von Leistungen wird durch die sehr heterogenen Lernvoraussetzungen in inklusiven Lerngruppen jedoch in Frage gestellt, da sozial vergleichende Bewertungen unter Umständen nicht die individuelle Anstrengung sowie den persönlichen Lernfortschritt berücksichtigen. Wie unterschiedlich Leistungsvoraussetzungen ausfallen können, verdeutlicht ein Blick auf die Einteilung sonderpädagogischer Förderschwerpunkte. Im Schuljahr 2018/2019 wurde 544.640 Lernenden der Primar- und Sekundarstufe I in Deutschland ein sonderpädagogischer Förderbedarf zugeschrieben, was einem Anteil von 7,4 Prozent entspricht (Hollenbach-Biele & Klemm, 2020). Die Gruppe der Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf setzt sich wiederum aus sehr heterogenen Subgruppen zusammen, wobei folgende offizielle Förderschwerpunkte differenziert werden: Lernen (35 % aller Lernenden mit sonderpädagogischem Förderbedarf), emotionale und soziale Entwicklung (18 %), geistige Entwicklung (17 %), Sprache (10 %), körperliche und motorische Entwicklung (7 %), Hören (4 %) und Sehen (2 %) (Hollenbach-Biele & Klemm, 2020). Lernende, deren Unterstützungsbedarf keinem spezifischen Förderschwerpunkt entspricht, werden der Kategorie *übergreifend* zugeordnet (7 %). Weiterhin werden Schülerinnen und Schüler, die aufgrund einer längerfristigen und schwerwiegenden Erkrankung im Krankenhaus oder zu Hause unterrichtet werden, in offiziellen Statistiken zur sonderpädagogischen Förderung aufgeführt (KMK, 2022). Bezüglich der genannten sonderpädagogischen Förderschwerpunkte und deren anteilsmäßigen Verteilung sei angemerkt, dass sich die Regularien und Vorgehensweisen zur Feststellung eines sonderpädagogischen Förderbedarfs je nach Bundesland deutlich unterscheiden können (Klemm, 2018, S. 8). Der Begriff des sonderpädagogischen Förderbedarfs ist dementsprechend – wie auch der Behinderungsbegriff – als biopsychosoziales Konstrukt zu verstehen. Dieses umfasst sowohl biologische Merkmale und psychosoziale Aufwachsens-Bedingungen des betroffenen Kindes als auch soziale Normvorstellungen, z.B. bezüglich des Verhaltens und der Leistung eines Kindes in einem bestimmten Alter oder auch in Hinblick auf die Unterschiede einer allgemeinen schulischen und sonderpädagogischen Förderung.

Alternative Formen der Leistungsbewertung

Die Begriffe der alternativen bzw. der neuen Formen schulischer Leistungsbewertung werden in der Literatur und auch im nachfolgenden Text synonym gebraucht. Ebenfalls werden die Termini Leistungsbeurteilung und Leistungsbewertung synonym benutzt. Leichte Unterschiede in der Begriffsverwendung werden u.a. bei Bohl (2019, S. 416) aufgegriffen, sind im Rahmen der nachfolgenden Ausführungen jedoch nicht relevant.

Alternative Formen der schulischen Leistungsbewertung berufen sich auf ein konstruktivistisches Lehr-Lern-Verständnis, welches die Eigenaktivität und Selbststeuerung des Lernenden hervorhebt (Anderson, 1998; Maclellan, 2004). In der Literatur wird insbesondere die pädagogische Funktion alternativer Bewertungsformen hervorgehoben (Fischer, 2012; Middendorf, 2012). Sie sollen der Optimierung individueller Lernprozesse und des methodisch-didaktischen Arrangements des Unterrichts dienen (Schmidinger, 2012). Zudem stellen das selbstgesteuerte Monitoring von Lernprozessen und die Selbstbewertung durch die Lernenden zentrale Elemente dar. Die qualitative Beschaffenheit einer Leistung und ihre Entstehung sollen inhaltlich beschrieben und unter Anwendung indivi-

dualisierter Bewertungsmaßstäbe gewürdigt werden, anstatt die entsprechenden Aspekte auf einen wenig informationshaltigen, abstrakten Zahlenwert zu reduzieren (Beutel, 2012; Winter, 2004). Grunder und Bohl (2004) knüpfen ihre Definition neuer Formen der Leistungsbeurteilung zudem an einen erweiterten Lernbegriff. Dieser umfasse neben dem fachlich-inhaltlichen Faktenwissen auch methodisch-strategische, sozial-kommunikative und persönlichkeitsbezogene Aspekte. Begründet werden alternative Formen der Leistungsbewertung z.B. über die angenommene Verringerung von Leistungsdruck, die vermutete Motivationssteigerung sowie die angestrebte Fokussierung auf inhaltliche Rückmeldungen und Lernhinweise (Bohl, 2019). In der Literatur werden alternative Bewertungsformen und ein unterrichtsbegleitendes formatives Assessment zudem in den übergreifenden Kontext einer sogenannten *neuen Lernkultur* eingebettet (Schmidinger et al., 2016, S. 69). Typischerweise gelten durch die Lehrperson verfasste verbale Beurteilungen oder Lernentwicklungsberichte sowie von den Lernenden erstellte Lerntagebücher, Portfolios, gegenseitige Peer-Bewertungen unter Schülerinnen und Schülern und ihre Selbstbewertungen, z.B. über Kompetenzraster, als alternative Formen der Leistungsbeurteilung (Middendorf, 2012; Schmidinger et al., 2016).

Das im Rahmen der empirischen Studie eingehender betrachtete Portfolio bezeichnet eine ausgewählte Sammlung von Arbeiten eines Lernenden, in denen sich individuelle Anstrengungen, Fortschritte und Leistungen in einem oder mehreren Lernbereichen widerspiegeln (Schmidt, 2020). Die Auswahl der Inhalte erfolgt unter Beteiligung des betreffenden Kindes. Die Beurteilungskriterien sollen transparent dargelegt und idealerweise gemeinsam entwickelt, die Selbstreflexion der Lernenden angeregt werden (Brunner et al., 2008). Differenzieren lassen sich vielfältige Formen des Portfolios, wie z.B. das Projekt-, Kurs-, Bewerbungs- und Lern-Entwicklungs-Portfolio sowie standardisierte Formen des Portfolios (Häcker, 2011; Schmidt, 2020).

Unter einem Kompetenzraster ist eine Tabelle zu verstehen, die Wissen und Fertigkeiten inhaltsspezifisch beschreibt und in unterschiedliche Niveaustufen aufgliedert: Eine Dimension spezifiziert Inhaltsbereiche in einem bestimmten Fachgebiet. Die zweite Dimension konkretisiert, welche Kompetenzniveaus Lernende in den thematischen Bereichen erlangt haben (Schmidt, 2020). Beispielhaft für den Einsatz des Kompetenzrasters im inklusiven Unterricht beschreiben Sasse und Schulzeck (2014, 2021) die sogenannte *Differenzierungsmatrix*. Angelehnt an die Struktur-Niveau-Theorie schulischen Lernens nach Kutzer (1982) beschreibt eine entsprechende Matrix zum einen die sachlogische Struktur eines Lerngegenstandes und zum anderen unterschiedliche kognitive Niveaustufen. Hierbei kann die Auseinandersetzung mit einem Inhalt auf einer basalen, konkret handlungsbezogenen Ebene erfolgen oder auch das Niveau abstrakter Denkoperationen annehmen. Über transparente Leistungskriterien im Kompetenzraster könnten Schülerinnen und Schüler ihren Lernprozess nachvollziehen. Zudem sei eine Quantifizierung erreichter Kompetenzstufen über ein Punktesystem denkbar (Sasse & Schulzeck, 2014).

Empirische Validierung alternativer Formen der Leistungsbewertung

Die nachfolgend berichteten Reviews, Metanalysen und Primärstudien geben Aufschluss über die messtheoretischen Gütekriterien und die Effekte alternativer Bewertungsformen (Andrade & Valtcheva, 2009; Double et al., 2020; Wiliam, 2011; Yan et al., 2021).

Forschungsstand zu messtheoretischen Gütekriterien

Die Metaanalyse von Sanchez et al. (2017) untersucht u.a. die Zuverlässigkeit von Selbstbewertungen von Lernenden vom dritten bis zum zwölften Schuljahr. Sie konstatieren keine signifikanten Unterschiede zwischen den Lehrkraftbeurteilungen und Schülerinnen- und Schülerselbstbewertungen. Es zeige sich eine hohe Korrelation von durchschnittlich $r = .67$ (Sanchez et al., 2017). Brown und Harris (2013) hingegen berichten in ihrem Review, dass die Konsistenz von Schülerinnen- und Schülerselbstbewertungen und externen Bewertungen (z.B. Lehrkrafturteilen und Testleistungen) je nach Primärstudie deutlich schwanke. Die Korrelationen variierten von schwachen Zusammenhängen von .20 bis zu starken Assoziationen von .80, wobei nur in wenigen Untersuchungen Korrelationen von über .60 erreicht würden (Brown & Harris, 2013). Topping (2003) sowie Dochy et al. (1999) berichten in ihren Reviews u.a. von der Neigung leistungsstarker Lernender zur Selbstunterschätzung sowie der Tendenz leistungsschwacher Lernender zur Selbstüberschätzung im Vergleich zum Lehrkrafturteil. Zudem seien die Selbstbewertungen älterer und im Lernstoff weiter fortgeschrittener Schülerinnen und Schüler zumeist zutreffender als bei jungen Lernenden (Butler, 2018; Nagel & Lindsey, 2018). Andrade (2019) berichtet ferner von einer Tendenz von Jungen bzw. Männern zur Selbstüberschätzung im Rahmen der Selbstbewertung. Mädchen bzw. Frauen neigten hingegen zur Selbstunterschätzung eigener Leistungen. Über das Ausmaß der Übereinstimmung der jeweiligen Selbst- und Lehrkraftbewertungen entscheide u.a. die Art der Implementierung der Selbstbewertung (z.B. durch ein Training, Bewertungsraster). Empirischen Befunden zufolge können u.a. folgende Faktoren die Validität von Selbstbewertungen steigern (zusammenfassend: Andrade, 2019): Anleitungen und Feedback zu Prozessen der Selbstbewertung (Bol et al., 2012), die Vorgabe von Checklisten oder Bewertungsrastern mit Beurteilungskriterien (Panadero & Romero, 2014) sowie die Erfahrenheit der Schülerinnen und Schüler mit Methoden der Selbstbewertung (Nagel & Lindsay, 2018; Yilmaz, 2017).

Aktuelle Metaanalysen zum Thema Peer-Bewertungen berichten vergleichsweise hohe durchschnittliche Korrelationen von Peer- und Lehrkraftbeurteilungen von $r = .63$, bzw. $r = .68$ (Li et al., 2016; Sanchez et al., 2017). Hinsichtlich der Konsistenz von Peer- und Lehrkraftbewertungen zeigen sich in den einzelnen Primärstudien jedoch durchaus heterogene Befunde: So geben Falchikov und Goldfinch (2000), die 49 zwischen 1959 und 1999 publizierte Studien analysierten, eine mittlere Korrelation von Peer- und Lehrkraftbewertungen von $r = .69$ an sowie eine Varianz von $r = .14$ bis $r = .99$. Li et al. (2016), die 69 seit 1999 publizierte Studien in ihre Metaanalyse einbezogen, konstatieren ebenfalls variierende Übereinstimmungsmaße von $r = .33$ bis $r = .86$. In seinem narrativen Review berichtet Topping (2003, S. 69), dass über 70 Prozent der Studien im Hochschulkontext eine adäquate Reliabilität und Validität von Peerbewertungen konstatieren. Auch für den Primar- und Sekundarbereich hätten Studien zumeist eine gute Übereinstimmung von Peer-Bewertungen und Lehrkrafturteilen ergeben (Topping, 2003, S. 69). Folgende moderierende Faktoren gingen mit einer höheren Zuverlässigkeit von Peer-Bewertungen einher: Einsatz in Fortgeschrittenenkursen (Li et al., 2016), die gemeinsame Entwicklung und Diskussion der Bewertungskriterien mit den Lernenden (Li et al., 2016), die Verwendung von Bewertungsrastern oder Checklisten (Sanchez et al., 2017), ein Training und Monitoring der Bewertungsprozesse und diesbezügliches Feedback durch die Lehrperson (Liu & Li, 2014; Wanner & Palmer, 2018). Ferner nahmen die Erfahrenheit und kognitive Leistungsfähigkeit von Peer-Bewertenden Einfluss auf die Qualität von Peer-Bewertungen (van Zundert et al., 2010). In weiteren Einzelstudien zeigte

sich, dass Lehrkräfte strengere Beurteilungsstandards anlegten als Peer-Bewertende (Chang et al., 2012) und dass auch unter Einsatz von Bewertungsrastern Leistungskriterien von Lehrkräften und Lernenden teilweise unterschiedlich interpretiert wurden (DeGrez et al., 2012).

In Bezug auf Lernportfolios sind allgemeine Aussagen zur Reliabilität und Validität nur in sehr eingeschränktem Maße möglich (Meeus et al., 2009). Zum einen werden Portfolios in sehr unterschiedlichen Formaten und Kontexten eingesetzt, wie z.B. im Schulkontext (Chang & Wu, 2012) oder in der universitären Bildung von Medizinstudierenden (Driessen, 2008), angehenden Zahnärzten (Gadbury-Amyot et al., 2014) und Lehramtsstudierenden (Binh, 2021; Meeus et al., 2008). Zum anderen erheben insbesondere formativ genutzte Portfolios den Anspruch, einen authentischen Eindruck des subjektiven Lernprozesses zu vermitteln und Prozesse der Selbstreflexion zu befördern. Driessen et al. (2005) verweisen darauf, dass eine Standardisierung von Portfolios, z.B. im Rahmen einer vorgegebenen Strukturierung und vorgeschriebener Inhalte, die Reliabilität der Portfolio-Bewertung steigern mag. Jedoch könnte die persönlich-authentische Selbstreflexion unter einer entsprechenden Standardisierung leiden. Die Autoren plädieren daher für eine Verwendung qualitativer Kriterien, um die Güte der Portfoliobewertung zu ermessen, z.B. über eine kommunikative Validierung durch mehrere Bewertende und die Diskussion von Bewertungen zwischen Dozierenden und Studierenden (Driessen et al., 2005). Bei der Quantifizierung klassischer testtheoretischer Gütekriterien verweisen einzelne Studien auf eine vergleichsweise hohe Interrater-Reliabilität bei der Portfoliobewertung (Heller et al., 1998; Herman, et al., 1993). Andere Untersuchungen konstatieren erhebliche Schwankungen der Interrater-Reliabilität, insbesondere wenn eine exakte Übereinstimmung durch unabhängige Beurteilende als Kriterium zugrunde gelegt wird (Baume & Yorke, 2002; Nystrand et al., 1993; Tochel et al., 2009). Gadbury-Amyot et al. (2014) berichten u.a. von schwachen, aber signifikanten Zusammenhängen der Portfoliobewertung von 73 Studierenden der Zahnmedizin mit den Noten staatlicher Abschlussprüfungen (r = .32 und r = .27). Chang und Wu (2012) ermittelten gute Übereinstimmungen der Bewertung eines elektronischen Portfolios mit den kursinternen Abschlussnoten von 72 Lernenden einer High School sowie eine akzeptable Interrater-Reliabilität unter Verwendung eines standardisierten Bewertungsschemas. Driessen et al. (2006) konstatieren ferner moderate bis sehr gute Interrater-Reliabilitäten (r = .46 bis r = .87) bei der Bewertung von 40 durch Medizinstudierende eingereichten Portfolios. Die Verwendung eines einheitlichen Bewertungsschemas mit klaren und expliziten Standards, die Vertrautheit mit den entsprechenden Bewertungskriterien sowie ein Training der Beurteilenden erhöhen die Interrater-Reliabilität (Favier et al., 2019; Tochel et al. 2009).

Abschließend sei darauf verwiesen, dass Aspekte der Reliabilität und Validität alternativer Bewertungsformen in der Literatur durchaus kritisch hinterfragt werden (u.a. Andrade, 2019; Favier et al., 2019; Maclellan, 2004). Unter Umständen würden randständige Aspekte in Bezug auf die zu messende Leistung erfasst und relevante Aspekte übersehen (Maclellan, 2004). Beispielsweise ist es denkbar, dass im naturwissenschaftlichen Unterricht eingesetzte Portfolios eher die verbale Ausdrucksfähigkeit von Schülerinnen und Schülern widerspiegeln und weniger ihr fachbezogenes Wissen oder logisches Denken. Lawrenz et al. (2001) verweisen in diesem Kontext darauf, dass Lernende mit unterschiedlichen ethnischen Hintergründen unter variablen Beurteilungsformen (z.B. Multiple-Choice-Aufgaben, frei zu beantwortende oder handlungsorientiert praktische Tests) variable Ergebnisse erzielen. Ferner deuten inhaltsanalytische Untersuchungen darauf hin, dass der Portfolio-Ansatz auch im universitären Kontext nicht automatisch

mit einer validen Dokumentation selbstregulierter Lernprozesse einhergeht (van der Gulden et al., 2020).

Forschungsstand zu lernförderlichen und motivationalen Effekten

Allgemein sind positive Effekte der Selbstbewertung auf das akademische Lernen von Schülerinnen und Schülern und Studierenden zu konstatieren, insbesondere wenn die Selbstbewertung zur formativen Optimierung des weiteren Lernprozesses eingesetzt wird (Andrade, 2019; Brown et al., 2015; Brown & Harris, 2013). Eine aktuelle Metaanalyse von Yan et al. (2021) ermittelte eine mittlere Effektstärke von g = .46 in Hinblick auf die leistungsförderliche Wirkung von Selbstbewertungen auf das akademische Lernen im Hochschulkontext. Moderiert wurde der Effekt durch explizite Rückmeldungen zu den entsprechenden Selbstbewertungen durch Mitstudierende oder Dozierende: Unter Einsatz externen Feedbacks zur jeweils vorgenommenen Selbstbewertung fiel die Effektstärke wesentlich höher aus als ohne eine entsprechende Rückmeldung von außen (g = .66 versus g = .21; Yan et al., 2021). In der Metaanalyse von Li und Zhang (2020) ergab sich ein moderater Zusammenhang (r = .47, p < .01) zwischen dem Einsatz von Selbstbewertungen und dem Lernfortschritt von Schülerinnen und Schülern im sprachlichen Bereich. Wie effektiv die Selbstbewertungen in den jeweiligen Primärstudien ausfielen, war v.a. von den zur Selbstbewertung eingesetzten Bewertungsinstrumenten, der Formulierung klarer Beurteilungskriterien und einem Training der Lernenden abhängig. Sanchez et al. (2017) berichten in ihrer Metaanalyse von besseren Testleistungen von Lernenden, die sich im Vorfeld selbst bewerteten (g = 0.34). Graham et al. (2015) ermitteln eine durchschnittliche gewichtete Effektstärke von 0.62 in Bezug auf die Förderung von schriftsprachlichen Leistungen von Schülerinnen und Schülern durch Maßnahmen der Selbstbewertung. Weiterhin verweisen empirische Studien auf ein höheres Lernengagement, vertiefte Lernprozesse und eine erhöhte Selbstwirksamkeitsüberzeugung, die mit Selbstbewertungen in Bezug stehen (zusammenfassend: Topping, 2003). Ferner wird von einer Förderung des selbstregulierten Lernens (Wang, 2017) sowie des kritischen Denkens (Siow, 2015) berichtet. Hierbei ist allerdings zu berücksichtigen, dass Selbstbewertungen in der Regel keine isolierte pädagogische Intervention darstellen. Oft gehen Aspekte wie die selbstregulierte Planung und Sequenzierung von Lernprozessen sowie die Selbstüberwachung mit der selbstständigen Bewertung eigener Leistungen einher (Topping, 2003, S. 64).

In ihrer Metaanalyse untersuchten Double et al. (2020) die Effekte von Peerbewertungen auf akademische Leistungen. Hierbei beziehen sie 54 Studien mit experimentellem oder quasiexperimentellem Kontrollgruppen-Design ein. In der Zusammenschau ergeben sich schwache bis mittlere Effekte der Peerbewertung in Hinblick auf die akademische Leistung von Lernenden unterschiedlichen Alters. Die Vorteile gegenüber einem kompletten Verzicht auf Bewertungen fallen hierbei am deutlichsten aus (g = 0.31, p < .004). Auch gegenüber der Bewertung durch die Lehrkraft zeichnen sich leichte lernförderliche Effekte ab (g = 0.28, p < .007). Keine signifikanten Unterschiede in Hinblick auf lernförderliche Wirkungen bestehen im Vergleich zur Selbstbewertung (Double et al., 2020). Weiterhin berichten narrative Reviews von lernförderlichen Effekten der Peerbewertung (u.a. Topping, 2003; van Zundert et al., 2010). Kritisch anzumerken ist, dass nur wenige empirische Studien zur Peerbewertung ein Kontrollgruppendesign und objektivierbare Indikatoren des Lernfortschritts aufweisen. Häufig stünden Selbstauskünfte und Wahrnehmungen von Lernenden oder Lehrkräften im Fokus der Betrachtungen (Double et al., 2020). Als moderierende Einflüsse, welche die Wirksamkeit der Peerbewertung steigern können, wurden in Einzelstudien u.a. folgende Variablen identifiziert:

Anleitung und Training von Prozessen der Peerbewertung (Panadero & Jonsson, 2013; van Zundert et al., 2010; Wanner & Palmer, 2018), Nutzung von Bewertungsrastern (Panadero et al., 2013) und die Erfahrenheit von Lernenden mit Formen der Peerbewertung (Smith et al., 2002; van Zundert et al., 2010). Die Metaanalyse von Double et al. (2020) konnte allerdings keine Signifikanz der diskutierten Moderatoren nachweisen. Li et al. (2020) sprechen von einer heterogenen Befundlage in Hinblick auf die Lernwirksamkeit der Peer-Bewertung. Ihre Metaanalyse stützt sich auf 58 Primärstudien und konstatiert eine durchschnittliche Leistungssteigerung um .29 Standardabweichungen infolge der Peer-Bewertung.

Im Zusammenhang mit der Portfoliobewertung berichten empirische Studien von: einem aktiveren Lernverhalten (Isnawati et al., 2021), der Förderung des kritischen Denkens (Muho & Leka, 2021; Tiara Linanti et al., 2021), einer vermehrten Selbstregulation beim Lernen (Chang et al., 2018; Mak & Wong, 2018), einer Erhöhung der Schreibkompetenz (Biglari et al., 2021) und dem erfolgreichen Erlernen von Fremdsprachen (Nezakatgoo, 2011; Segaran & Hasim, 2021). Händel et al. (2018) untersuchten den Einsatz eines zur Reflexion anregenden elektronischen Portfolio zur Prüfungsvorbereitung in einer Stichprobe von insgesamt 1469 Studierenden. Die Studierenden, die das elektronische Portfolio nutzten, schnitten in der Abschlussklausur signifikant besser ab als die Kontrollgruppe ohne entsprechendes Portfolio. Zudem erwies sich die Zeit, die auf die Bearbeitung des elektronischen Portfolios verwendet wurde, als ein signifikanter Prädiktor der Abschlussnote (Händel et al., 2018). Die Potenziale des Portfolioansatzes werden u.a. für den Bereich der universitären und beruflichen Bildung und insbesondere in Hinblick auf die Anregung einer kritischen Selbstreflexion hervorgehoben (Binh, 2021; Driessen, 2008). Bei Lernenden mit sonderpädagogischem Förderbedarf sind kognitive und selbstreflexive Kompetenz jedoch individuell abzuschätzen, um eine sinnvolle und niveauangepasste Portfolioarbeit zu initiieren (Dharma & Hermanto, 2019).

Weiterführende Erkenntnisse zum Einsatz alternativer Bewertungsformen in einer heterogenen Lerngruppen liefert u.a. ein zweijähriges Modellprojekt zur Einführung von Kompetenzportfolios an fünf nordrhein-westfälischen Waldorfschulen (Brater et al., 2010). Diesbezüglich wird kritisch resümiert, dass lediglich einzelne Lernendenarbeiten auf eine erhöhte Selbstreflexivität und ein gesteigertes Verständnis für die eigenen Lernprozesse verwiesen. Insbesondere sprachlich und kognitiv weniger leistungsstarke Schülerinnen und Schüler profitierten oft nicht von der Arbeit mit Portfolios (Brater et al., 2010, S. 199). Laut Grittner (2009) betrachteten Grundschullehrkräfte das Portfolio prinzipiell als aussagekräftig und informativ in Bezug auf die Fähigkeiten und das Arbeitsverhalten von Lernenden. Jedoch monierten sie den Zeitaufwand der Portfolioarbeit. Arnold et al. (2000) begleiteten einen Schulversuch zu alternativen Formen der Leistungsbeurteilung. Die diesbezüglichen Rückmeldungen fielen variabel aus: So berichtete eine Schule von einer erhöhten mündlichen Beteiligung und mehr Engagement insbesondere von Seiten leistungsschwächerer Schülerinnen und Schüler. Eine Verbesserung schulischer Leistungen sei allerdings nicht beobachtet worden. An anderen Schulen seien vor allem leistungsschwächere Lernende mit der selbstständigen Reflexion ihres Lernprozesses überfordert gewesen (Arnold et al., 2000).

Hinsichtlich alternativer Formen der Leistungsbeurteilung wurde in Deutschland insbesondere auf mögliche differenzielle Wirkungen der Verbalbeurteilung im Gegensatz zur Notengebung fokussiert (Jachmann, 2003, S. 68f.). In einer Untersuchung unter 241 Grundschulkindern fanden Wagner und Valtin (2003) jedoch keine belastbaren Belege für einen leistungs- und entwicklungsförderlichen Effekt von verbalen Beurteilungen im Vergleich zu Noten. Weiterhin wiesen Textanalysen auf Mängel

in der praktischen Umsetzung der Verbalbeurteilung hin: Differenzierte und informative Rückmeldungen, Ermutigungen der Lernenden sowie dezidierte förderdiagnostische Hinweise seien hierin häufig nicht enthalten (Schmude, 2001; Valtin, 2012). In der Hamburger Lernausgangs-Untersuchung (LAU) fand ebenfalls ein Vergleich zwischen Grundschulkindern mit Verbalbeurteilung und Gleichaltrigen mit Ziffernbenotung statt (Lehmann et al., 1997, 1999). Hierbei führte der Verzicht auf Noten zu keinen Leistungseinbußen. Viertklässlerinnen und -klässler ohne Noten schnitten in den standardisierten Leistungstests sogar etwas besser ab und wiesen ein etwas höheres Selbstkonzept auf (Lehmann et al., 1997, 1999).

In der Untersuchung von Jachmann (2003) schätzten Sechstklässlerinnen und -klässler mit Ziffernbenotung die Lernkultur in ihrer Klasse tendenziell weniger positiv ein und gaben eine etwas stärkere Schulunlust an als Lernende der sechsten Jahrgangsstufe, die Verbalbeurteilungen erhielten. Ein Zusammenhang zwischen der an der jeweiligen Schule praktizierten Beurteilungsform und der Schulangst der Schülerinnen und Schüler konnte hingegen nicht bestätigt werden (Jachmann, 2003).

Forschungsstand zur formativen Leistungsbewertung

Die formative Leistungsbewertung stützt sich auf eine solide empirische Basis (Stiggins et al., 2007; Wiliam, 2011). Dies gilt sowohl in Bezug auf die Leistungszuwächse der Schülerinnen und Schüler in standardisierten Leistungstests als auch ihre Lernmotivation. Die Quantifizierung leistungsförderlicher Effekte des formativen Assessments in Reviews variiert allerdings von $d = 0.20$ bis 0.70, je nachdem welche Primärstudien einbezogen werden (Black & Wiliam, 1998; Kingston & Nash, 2011; zusammenfassend: Schmidt, 2020). Insgesamt handelt es sich bei der formativen Leistungsbewertung um eines der überzeugendsten unterrichtlichen Rahmenkonzepte (zusammenfassend: Schütze et al., 2018). Auch Mitchell (2014), der in seinem Forschungsreview effektive Strategien für den Unterricht in inklusiven und sonderpädagogischen Lernsettings betrachtet, attestiert dem formativen Assessment in Verbindung mit Feedback eine starke empirische Evidenz entsprechend einer Effektstärke von mindestens $d = 0.70$.

McLaughlin und Yan (2017) verweisen in ihrem Literaturreview über 75 zwischen 1998 und 2016 publizierten Studien ebenfalls auf die positiven Effekte onlinebasierter formativer Assessmentformen auf die Leistung und Selbstregulation von Lernenden. Gikandi et al. (2011) betrachten in ihrem Review ebenfalls die Effekte digitaler Formen der formativen Leistungsbeurteilung unter aktiver Beteiligung der Lernenden (z.B. selbstständig durchführbare Tests, Diskussionsforen und webbasierte Portfolios). Nach Gikandi et al. (2011) tragen die beschriebenen alternativen Formen der Leistungsbewertung zu einem erhöhten Lernengagement bei, wobei insbesondere ein interaktives Feedback lernförderlich wirkt.

Rakoczy (2012) beschreibt eine mit 329 deutschen Realschülerinnen und -schülern durchgeführte Laborstudie zu den Effekten der formativen Leistungsbeurteilung. Unter anderem wurde eine sozial vergleichende und eine auf den individuellen Lösungsprozess bezogene Rückmeldebedingung nach einem Mathematiktest realisiert. Das lösungsprozessbezogene Feedback ging hierbei auf die Stärken und Schwächen des jeweiligen Lernenden ein und explizierte Strategien zur Leistungsverbesserung. Es trug in besonderem Maße dazu bei, dass die Lernenden besser wussten, wie sie ihren weiteren Lernprozess produktiv gestalten konnten. Zudem ergab sich ein positiver Zusammenhang zwischen einer lösungsprozessbezogenen Rückmeldung und einem erhöhten Leistungszuwachs sowie gesteigertem Aufgabeninteresse (Harks et al., 2014).

Das IGEL-Projekt (Individuelle Förderung und adaptive Lern-Gelegenheiten in der Grundschule) verweist ebenfalls auf die

lernförderlichen Effekte einer unterrichtsbegleitenden formativen Bewertung (Decristan et al., 2015). Realisiert wurde u.a. eine quasiexperimentelle Teiluntersuchung mit Kontrollgruppe unter Einbezug von 551 Kindern der dritten Jahrgangsstufe und 28 Lehrpersonen. Die in der Experimentalgruppe unterrichtenden Lehrpersonen wurden instruiert, den Schülerinnen und Schülern Rückmeldungen über halbstrukturierte Feedbackbögen zu geben, in denen der aktuelle Lernstand eingestuft und Hinweise für das weitere Lernen vermittelt wurden (Decristan et al., 2015). Im Vergleich zur Kontrollgruppe ohne eine entsprechende formative Leistungsbewertung schnitten die Kinder der Experimentalgruppe in abschließenden Leistungstests erfolgreicher ab.

Im Rahmen des Modellversuchs KOMPASS (Kompetenz aus Stärke und Selbstbewusstsein) wurden ebenfalls neue Formen einer formativen und auf Schülerinnen- und Schüler-Partizipation ausgerichteten Leistungsbeurteilung erprobt, wie z.B. der Einsatz von Kompetenzrastern, Checklisten, Portfolios und individuellen Rückmeldungen (Scheunpflug et al., 2012). An der Längsschnittuntersuchung mit Kontrollgruppendesign waren zwölf bayerische Realschulen mit insgesamt über 3600 Schülerinnen und Schülern sowie 800 Lehrkräfte beteiligt. Die Evaluation des realisierten Modellversuchs spricht für günstige Effekte, u.a. in Hinblick auf die Lernmotivation, das schulische Interesse und Selbstkonzept der Lernenden. Im Vergleich zu Gleichaltrigen an Kontrollschulen gaben Lernende an Interventionsschulen in verstärktem Maße an, ihre Fähigkeiten für veränderbar zu halten und konstruktiv mit Misserfolgen umzugehen (Scheunpflug et al., 2012). Weitere Studien verweisen auf eine höhere Lernfreude, intrinsische Lernmotivation, Kompetenzerleben und Autonomie beim Lernen in Bezug auf Schülerinnen und Schüler mit formativ angelegter Leistungsbewertung (Dori, 2003; Maslovaty & Kuzi, 2002; Rakoczy et al., 2008; Smit, 2009). Bürgermeister (2014) berichtet u.a. von einer erhöhten Anstrengungsbereitschaft und Motivation für den Mathematikunterricht, wenn Schülerinnen und Schüler partizipativ in den Beurteilungsprozess mit einbezogen werden.

Insgesamt ist zu berücksichtigen, dass das formative Assessment ein uneinheitliches Konzept darstellt. In unterschiedlichen Studien werden differenzierte Teilaspekte des Konstrukts variabel umgesetzt und unterschiedlich operationalisiert (Taras, 2010). Empirisch validierte Komponenten vieler Verfahren der formativen Leistungsbewertung sind v.a. Selbstbewertung (Andrade & Valtcheva, 2009; Dochy et al., 1999; Tochel et al., 2009) und Feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006). Zudem ist es erfolgskritisch, dass Erkenntnisse der formativen Bewertung auch tatsächlich didaktische Adaptionen im Unterricht nach sich ziehen (McMillan et al., 2013).

Weiterhin existieren Einzelstudien, die keine signifikanten Effekte des formativen Assessments finden (Yin et al., 2008). Zudem wird darauf verwiesen, dass einige Lehrpersonen die Nutzung einer unterrichtsbegleitenden Leistungserfassung für z.B. Rückmeldegespräche und eine diesbezügliche Adaption des Lernangebotes als zu aufwendig und unterrichtspraktisch kaum realisierbar einstufen (Hebbecker et al., 2020). Ferner werfen Falkenberg et al. (2017, S. 317) einen kritischen Blick auf die schulische Leistungsbeurteilung in Schweden, die „zwischen formativem Anspruch und summativer Notwendigkeit“ stehe. Sie berichten, dass in der Beurteilungspraxis die bis zur sechsten Jahrgangsstufe rein formativ angelegten Lernstandsrückmeldungen teilweise durch summative Beurteilungselemente überlagert würden. Dies führe zu widersprüchlichen Anforderungen an die Lehrkräfte, welche diese als solche erkennen und versuchen auszubalancieren. In Bezug auf die Lernenden bringe eine formative Leistungsbewertung auch andere Effekte als die primär intendierten mit sich, nämlich das Gefühl eines ständigen Geprüftwerdens (Falkenberg et al., 2017).

Forschungsstand zur Einschätzung alternativer Formen der Leistungsbewertung

Empirische Studien sprechen dafür, dass Lehrende, Lernende und Eltern mit alternativen Formen der Leistungsbeurteilung, wie z.B. Verbalbeurteilungen, dem Portfolio oder der Schülerinnen- und Schülerselbstbewertung, durchaus Vorteile verbinden (Andrade & Du, 2007). Dies resultiert jedoch nicht in einer Ablehnung traditioneller Formen der Leitungsbeurteilung. Die gängige Benotungspraxis wird in der Regel nicht hinterfragt und gilt trotz empirisch erwiesener, messtheoretischer Mängel zumeist als objektiv (Jachmann, 2003; Lau & Lübeck, 2021; Valtin, 2012).

Jachmann (2003) berichtet, dass die Vorstellung einer Schule ohne Noten von Lernenden mehrheitlich abgelehnt wurde. Wenn sich die befragten Schülerinnen und Schüler zwischen Berichts-, Noten- oder Notenzeugnissen mit Kommentarbogen entscheiden dürften, so gab etwa die Hälfte (53 Prozent) an, das Notenzeugnis mit Kommentarbogen zu präferieren. Knapp 40 Prozent bevorzugten das Notenzeugnis und nur circa sieben Prozent das Berichtszeugnis (Jachmann, 2003). Prengel (2007) befragte Kinder der vierten bis sechsten Klasse der Montessori-Gesamtschule Potsdam zu ihren Erfahrungen mit neu eingeführten Pensenbüchern. Bei knapp der Hälfte der Kinder dominierte eine positive Sicht. Die anderen Schülerinnen und Schüler äußerten sich hingegen ambivalent oder ablehnend. Negativ wurde angemerkt, dass die Konfrontation mit dem zu erreichenden Pensum zu Gefühlen der Überforderung führen könne (Prengel, 2007).

Eltern hinterfragten die etablierte Praxis traditioneller Notengebung zumeist nicht (Dzelili, 2009; Jachmann, 2003). Noten würden als eindeutig und aussagekräftig eingestuft und insbesondere nach den ersten Grundschuljahren durch Eltern erwünscht (Dzelili, 2009; Jachmann, 2003).

Lehrpersonen monieren u.a. den hohen Arbeitsaufwand zur Erstellung von Verbalbeurteilungen sowie zur Etablierung alternativer Formen der Leistungsbewertung (Arnold et al., 2000; Brater et al., 2010; Grittner, 2009). In der Befragung von Jachmann (2003, S. 108) unter 637 Hamburger Lehrkräften sprachen sich 63 Prozent gegen eine Schule ohne Zensierung aus. Vor allem Lehrpersonen am Gymnasium, an der Haupt- und Realschule favorisierten die klassische Benotung. Eine kritische Haltung gegenüber der Ziffernbenotung nahmen v.a. Lehrkräfte an Grund- und Gesamtschulen ein. Auch internationale Studien zeigen, dass Noten und traditionelle Formen der Leistungsbewertung bei Lehrkräften in der Regel auf eine hohe Akzeptanz stoßen (Birgin & Baki, 2009; Dzelili, 2009; Watt, 2005).

Forschungsfragen

Nicht zuletzt durch die sukzessive Umsetzung schulischer Inklusion stellt sich zunehmend die Frage nach geeigneten Maßnahmen, welche die individuelle Leistung von Kindern mit und ohne besonderen Förderbedarf diagnostizieren und unterstützen können. Dabei kommen auch alternative, potentialorientierte Formen der Leistungsbeurteilung in den Blick. Um an diese beschriebenen Überlegungen im öffentlichen und im wissenschaftlichen Diskurs anzuknüpfen, lautet die zentrale Forschungsfrage der vorliegenden Studie: „Inwiefern sind ausgewählte alternative Formen der Leistungsbeurteilung wie Lernportfolios und Kompetenzraster bei Lehrkräften an Grundschulen bekannt und werden im Unterricht eingesetzt?“ Zusätzlich soll eruiert werden, inwiefern nach Einschätzung der Beteiligten der Einsatz der diversen alternativen Formen der Leistungsbeurteilung gerade in Bezug auf den Umgang mit Heterogenität und Inklusion in der Schule von Vorteil ist. Dementsprechend wurden folgende vier Bereiche eruiert:

1. Inwieweit sind den teilnehmenden Grundschullehrkräften verschiedene alternative Formen der Leistungsbeurteilung bekannt?
   Aufgrund des jahrzehntelangen Diskurses um mögliche Alternativen zur Zifferennote wurde ein vergleichsweise hoher Bekanntheitsgrad alternativer Formen der Leistungsbewertung angenommen.
2. Wie beurteilen die befragten Lehrpersonen die Eignung der einzelnen alternativen Bewertungsformen für die grundsätzliche Leistungsbeurteilung aller Schülerinnen und Schüler?
   Angenommen wurde, dass die befragten Lehrpersonen aufgrund von Studium oder Fortbildungen oder beidem mit einzelnen pädagogischen Vorteilen alternativer Formen der Leistungsbewertung vertraut sind, wie z.B. Motivationssteigerung und Hinleitung zu selbstreguliertem Lernen durch Verdeutlichung individueller Lernfortschritte (Bohl, 2019; Grunder & Bohl, 2004). Daher wurde insgesamt von einer recht hohen prinzipiellen Eignungseinschätzung ausgegangen. In Hinblick auf Selbstbewertungen könnte sich allerdings eine kritischere Eignungseinstufung abzeichnen, da Selbsteinschätzungen von jüngeren Schülerinnen und Schülern weniger valide sind (Dochy et al., 1999; Topping, 2003). Auch Gruppenarbeitsergebnisse könnten ggf. kritischer beurteilt werden, da hier der Beitrag einzelner Gruppenmitglieder nicht immer eindeutig zu identifizieren ist und sich ggf. ein Trittbrettfahrereffekt einstellen kann.
3. Wie beurteilen die befragten Lehrpersonen die Eignung der einzelnen alternativen Bewertungsformen für die Leistungsbeurteilung von Schülerinnen und Schülern mit besonderem Förderbedarf? Bestehen gruppenbezogene Unterschiede in den jeweiligen Eignungseinstufungen?
   Angesichts der Potenziale alternativer Bewertungsformen in Hinblick auf eine Würdigung von Lernanstrengungen und die Anwendung einer individuellen Bezugsnorm wird von einer eher optimistischen Eignungseinschätzung ausgegangen (Beutel, 2012; Grunder & Bohl, 2004). Aufgrund der kognitiven und sprachlichen Anforderungen einzelner alternativer Bewertungsformen könnte die Eignungseinschätzung in Bezug auf Lernende mit sonderpädagogischen Förderbedarf jedoch etwas pessimistischer ausfallen als bei Schülerinnen und Schülern ohne sonderpädagogischen Förderbedarf (Arnold et al., 2000; Brater et al., 2010).
4. Wie häufig setzen Lehrkräfte alternative und traditionelle Formen der Leistungsbeurteilung in ihrem Unterricht ein? Bestehen Zusammenhänge zwischen der Eignungseinschätzung und der angegebenen Einsatzhäufigkeit unterschiedlicher Formen der Leistungsbewertung?
   Es wird vermutet, dass klassische Formen der Leistungsbewertung, v.a. schriftliche Klassenarbeiten und Tests, nach wie vor häufiger eingesetzt werden als alternative Bewertungsformen, da die Etablierung entsprechender neuer Formen der Leistungsmessung als arbeitsaufwendig gilt (Arnold et al., 2000; Brater et al., 2010; Hebbecker et al., 2020). Weiterhin wird angenommen, dass Lehrpersonen, die spezifische Formen der alternativen Leistungsbewertung als geeigneter einstufen, die entsprechenden Beurteilungsformen häufiger einsetzen.

Methode

In einer querschnittlich angelegten Fragebogenstudie unter Lehrpersonen wurden am Ende des Schuljahres 2017 Daten zum Einsatz und zur Eignung alternativer Formen der Leistungsbewertung im inklusiven Grundschulunterricht erhoben. Die Erhebung erfolgte im Rahmen einer breiter angelegten Dissertation zu heilpädagogischen Interventionen in inklusiven Grundschulkontexten (Kaul, 2020).

Stichprobe

Es beteiligten sich 40 Lehrkräfte im Alter von 26-60 Jahren (M = 43,7; SD = 9,88) an 13 inklusiv arbeitenden Grundschulen in öffentlicher Trägerschaft in einer Großstadt in Nordrhein-Westfalen. Die Teilnahme sowohl der Schulen (rekrutiert über die Schulleitungen) als auch der Lehrkräfte erfolgte ausschließlich auf freiwilliger Basis, weshalb davon auszugehen ist, dass es sich um eher engagierte Lehrkräfte an aufgeschlossenen Schulen und somit eine vorselektierte Gruppe handelte, die sich an der Befragung beteiligte. Von den befragten Lehrpersonen waren N = 34 weiblich und N = 5 männlich. Bei einer Person fehlte die Angabe des Geschlechts. Zudem wurden die Lehrkräfte nach ihrer Berufserfahrung bzw. ihren Dienstjahren befragt. Bei der Angabe wurde das Referendariat mit einbezogen, wohingegen Erziehungsurlaub und andere längere Unterbrechungen nicht gezählt wurden. Die Spanne der Diensterfahrung der Befragten reichte von 2 bis 40 Jahren (M = 18,00; SD = 8,98).

Operationalisierung

Der eingesetzte Fragebogen ist eine Eigenkonstruktion und bestand aus geschlossenen Items. Erfasst wurden die Kenntnis und die eingeschätzte Eignung verschiedener Formen der Leistungsbewertung zur Leistungsbeurteilung von Schülerinnen und Schülern, insbesondere auch für Lernende mit sonderpädagogischem Förderbedarf. Darüber hinaus wurde die Häufigkeit des Einsatzes der verschiedenen traditionellen und alternativen Formen der Leistungsbewertung erhoben.

Die allgemeine Kenntnis unterschiedlicher Formen der alternativen Leistungsbewertung wurde über ein dichotomes Antwortformat erhoben (*kenne ich*, *kenne ich nicht*).

Die abzugebende Eignungseinschätzung bezog sich ausschließlich auf alternative Formen der Leistungsbewertung. Hierbei wurden alternative Bewertungsformen ausgewählt, die Lernende partizipativ in Bewertungsprozesse einbeziehen, selbstreguliertes Lernen fördern und die formative Gestaltung von Lernprozessen erleichtern. Konkret wurden folgende Einzelitems abgefragt: Lernportfolios, Kompetenzraster, Selbstbewertungen von Schülerinnen und Schülern (z.B. anhand von Lernjournalen, Checklisten oder Fragebögen) und Gruppenarbeitsergebnisse. Das Antwortformat war vierstufig und reichte von *sehr geeignet*, *eher geeignet*, *eher ungeeignet* bis *völlig ungeeignet*. Eine entsprechende Eignungseinstufung der genannten Bewertungsformen wurde sowohl für die grundlegende Leistungsbeurteilung aller Schülerinnen und Schüler als auch spezifisch für die Beurteilung von Lernenden mit sonderpädagogischem Förderbedarf erhoben. Abbildung 1 zeigt die Items zur Eignungseinschätzung alternativer Beurteilungsformen in Bezug auf Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf.

Die abzugebende Einschätzung in Bezug auf die Einsatzhäufigkeit unterschiedlicher Formen der Leistungsbewertung umfasste neben alternativen auch traditionelle Bewertungsformen, wie Klassenarbeiten und Tests. Die konkret einzustufenden Einzelitems lauteten: (schriftliche) Probearbeiten/Klassenarbeiten/Tests, (mündliche) Tests, (praktische) Tests (z.B. im Sport- oder im Kunstunterricht), Lernportfolios, Kompetenzraster, Selbstbewertungen von Schülerinnen und Schülern (z.B. anhand von Lernjournalen, Checklisten oder Fragebögen) und Gruppenarbeitsergebnisse. Das vierstufige Antwortformat reichte von *sehr häufig*, über *eher häufig* und *eher selten* bis *ganz selten/gar nicht*. Abbildung 2 zeigt die Items zur Häufigkeitseinschätzung.

Welche der genannten Formen ist Ihrer Meinung nach besonders geeignet, um der **Leistungsbeurteilung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf** gerecht zu werden?

	sehr geeignet	eher geeignet	eher ungeeignet	völlig ungeeignet
Lernportfolios	☐	☐	☐	☐
Kompetenzraster	☐	☐	☐	☐
Schülerinnen- und Schülerselbstbewertungen (z.B. anhand von Lernjournalen, Checklisten oder Fragebögen)	☐	☐	☐	☐
Gruppenarbeitsergebnisse	☐	☐	☐	☐

Abbildung 1
Items zur Eignungseinschätzung alternativer Beurteilungsformen in Bezug auf Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf

Wie häufig setzen Sie die verschiedenen Formen zur individuellen Leistungsbeurteilung in Ihrem Unterricht ein?

	sehr häufig	eher häufig	eher selten	ganz selten/ gar nicht
Schriftliche Probearbeiten / Klassenarbeiten / Tests	☐	☐	☐	☐
Mündliche Tests	☐	☐	☐	☐
Praktische Tests (z.B. im Sport- oder im Kunstunterricht)	☐	☐	☐	☐
Lernportfolios	☐	☐	☐	☐
Kompetenzraster	☐	☐	☐	☐
Schülerinnen- und Schülerselbstbewertungen (z.B. anhand von Lernjournalen, Checklisten oder Fragebögen)	☐	☐	☐	☐
Gruppenarbeitsergebnisse	☐	☐	☐	☐

Abbildung 2
Items zur Einsatzhäufigkeit unterschiedlicher traditioneller und alternativer Formen der Leistungsbeurteilung

Auswertung

Die Auswertung der Items erfolgte mithilfe der Statistiksoftware SPSS. Zur inferenzstatistischen Prüfung von Korrelationen und Mittelwertsunterschieden werden nachfolgend nicht-parametrische Testverfahren eingesetzt (Rang-Korrelations-Koeffizienten nach Spearman und Kendall, Friedman-Test, Vorzeichen-Test und Wilcoxon-Test). Diese nicht-parametrischen Verfahren wurden gewählt, da die vorliegenden Daten einer vergleichsweise kleinen Stichprobe entstammen, nicht normalverteilt sind (Kolmogorov-Smirnov-Test, $p > .01$; Shapiro-Wilks-Test, $p > .01$) und die Erhebung der Eignungseinschätzung über Einzelitems eher ein ordinales Skalenniveau nahelegt. Das Alphaniveau wurde auf $\alpha = .05$ festgelegt und bei den angestellten paarweisen Vergleichen gemäß der Bonferroni-Korrektur auf $\alpha = .01$ angepasst.

Ergebnisse

Bekanntheit alternativer Bewertungsformen

Die Mehrheit der befragten Lehrkräfte (80-90 Prozent) gab an, sämtliche der genannten alternativen Formen der Leistungsbewertung zu kennen. Abbildung 3 gibt einen Überblick über die jeweiligen Prozentsätze der Lehrpersonen, denen die entsprechenden alternativen Bewertungsformen bekannt waren.

Allgemeine Eignungseinstufung alternativer Bewertungsformen

Wie aus Abbildung 4 ersichtlich, stuften die befragten Lehrpersonen die abgefragten alternativen Bewertungsformen als überwiegend *sehr geeignet* oder *eher geeignet* für die Leistungsbeurteilung von Schülerinnen und Schülern im Allgemeinen ein. Lernportfolios und Kompetenzraster wurden mit einer kumulierten Zustimmungsrate von 82.5 bzw. 80 Prozent in den Kategorien *sehr geeignet* und *eher geeignet* als am geeignets-

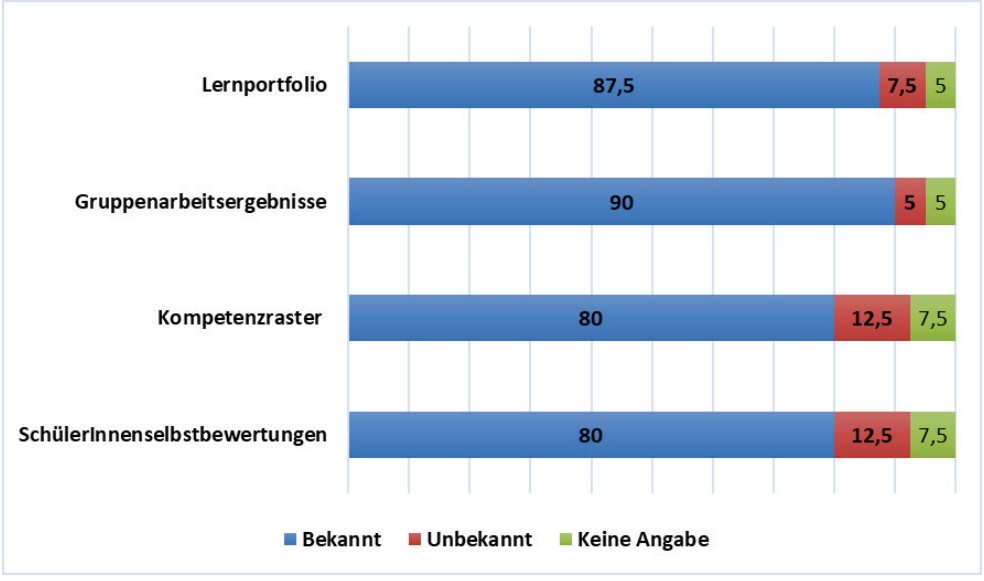

Abbildung 3
Bekanntheit alternativer Formen der Leistungsbewertung in Prozent (N = 40)

ten eingestuft. Die Eignungseinschätzung von Gruppenarbeitsergebnissen fiel am geringsten aus (kumulierte Zustimmungsrate von 62.5 Prozent in den Kategorien *sehr geeignet* und *eher geeignet*).

Eignungseinstufung alternativer Bewertungsformen bei sonderpädagogischem Förderbedarf

Abbildung 5 vermittelt einen Überblick über die jeweiligen Eignungseinstufungen alternativer Bewertungsformen für die Leistungsbeurteilung von Lernenden mit sonderpädagogischem Förderbedarf. Im Vergleich zu anderen alternativen Bewertungsformen wurden Lernportfolios als geeigneter für die Leistungsbeurteilung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf angesehen (kumulierte Zustimmungsrate von 67.5 Prozent in den Kategorien *sehr geeignet* und *eher geeignet*). Kompetenzraster und Schülerinnen- und Schülerselbstbewertungen wurden etwa von der Hälfte der Befragten als geeignet eingeschätzt, wohingegen Gruppenarbeitsergebnisse von den Lehrkräften insgesamt als eher ungeeignet eingestuft wurden (kumulierte Zustimmungsrate von 37.5 Prozent in den Kategorien *sehr geeignet* und *eher geeignet*).

Gruppenbezogene Unterschiede in den Eignungseinstufungen alternativer Bewertungsformen

In einem weiteren Schritt wurden die mittleren Lehrkraft-Bewertungen in Hinblick auf die Eignung variierender Formen der alternativen Leistungsbewertung für unterschiedliche Lernenden-Gruppen berechnet. Die Lehrpersonen konnten hierbei ein Rating auf einer vierstufigen Skala von *völlig ungeeignet* (1) bis *sehr geeignet* (4) vornehmen. Korrespondierend mit den berichteten Zustimmungsraten wurde das Lernportfolio sowohl für alle Schülerinnen und Schüler (M = 3.28; SD = 0.57) als auch explizit für Lernende mit sonderpädagogischem Förderbedarf (M = 3.48; SD = 0.63) als geeignet erachtet. In Bezug auf Gruppenarbeitsergebnisse fielen die eingeschätzten Eignungen für alle Schülerinnen und Schüler (M = 2.97; SD = 0.72) sowie für Lernende mit

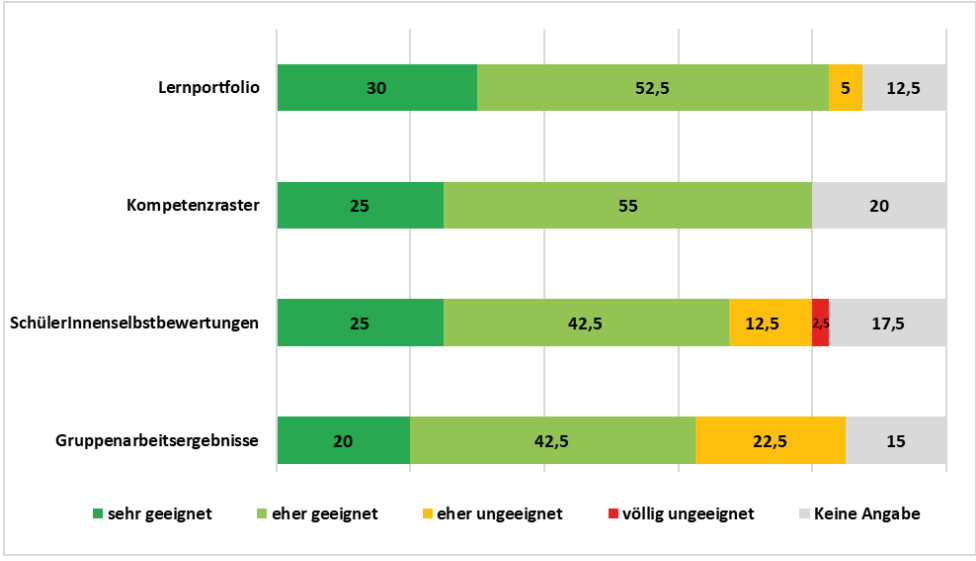

Abbildung 4
Eignungseinschätzung zu alternativen Formen der Leistungsbewertung für alle Schülerinnen und Schüler in Prozent (N = 40)

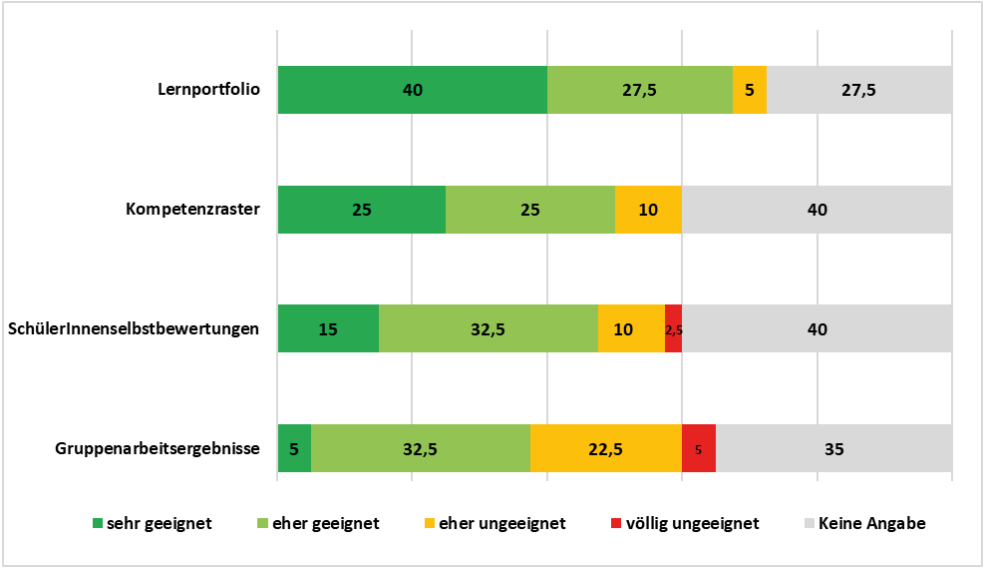

Abbildung 5
Eignungseinschätzung zu alternativen Formen der Leistungsbewertung für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf in Prozent (N = 40)

Tabelle 2
Mittlere Eignungseinstufung alternativer Leistungsbeurteilungsformen in Bezug auf unterschiedliche Schülerinnen- und Schülergruppen

Form der Leistungsbewertung	Alle Schülerinnen und Schüler		Schülerinnen und Schüler mit sonder-pädagogischem Förderbedarf		Mittelwerts-differenz
	M	*SD*	*M*	*SD*	
Lernportfolios	3.28	0.57	3.48	0.63	-0.20
Kompetenzraster	3.31	0.47	3.25	0.74	0.06
Schülerinnen- und Schülerselbstbewertungen	3.09	0.77	3.00	0.78	0.09
Gruppenarbeitsergebnisse	2.97	0.72	2.58	0.76	0.39***

Anmerkungen. *** $p < .001$ (gemäß Wilcoxon-Test und Vorzeichentest bei verbundenen Stichproben)

sonderpädagogischem Förderbedarf (M = 2.58; SD = 0.76) geringer aus und unterschieden sich deutlicher zwischen beiden Gruppen. Die deskriptiven Mittelwertunterschiede sind Tabelle 2 zu entnehmen.

Die inferenzstatistische Prüfung der zu konstatierenden deskriptiven Unterschiede in den Eignungseinstufungen alternativer Bewertungsformen erfolgte über den Friedman-Test für abhängige Stichproben. Der Friedman-Test verwies auf signifikante Unterschiede in den Bewertungen der erhobenen Beurteilungsformen (χ^2 [7] = 17.55, p < .014). Im Anschluss wurden paarweise Vergleiche der Eignungseinschätzungen alternativer Bewertungsformen für die generelle Leistungsbeurteilung aller Schülerinnen und Schüler sowie spezifisch für Lernende mit sonderpädagogischem Förderbedarf durchgeführt. Das Signifikanzniveau wurde mittels Bonferroni-Korrektur für vier wiederholte Testungen auf α = .01 angepasst. Hierbei zeigte sich, dass Gruppenarbeitsergebnisse zur Leistungsbeurteilung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf als signifikant weniger geeignet eingestuft wurden als zur allgemeinen Leistungsbeurteilung (Vorzeichentest: z = -3,18, p < .005; Wilcoxon-Text: z = -3,46, p < .005, r = .68, η^2 = .46). Die übrigen gruppenbezogenen Mittelwertsunterschiede erwiesen sich hingegen als nicht signifikant.

Einsatzhäufigkeit traditioneller und alternativer Bewertungsformen

Zwei traditionelle Formen der Leistungsbewertung wurden besonders häufig von den Lehrkräften eingesetzt: So entfiel auf schriftliche Probearbeiten/Klassenarbeiten/Tests eine kumulierte Zustimmungsrate von 80 Prozent in den Kategorien *sehr häufig* und *eher häufig* sowie auf praktische Tests, z.B. im Sport oder im Kunstunterricht, eine kumulierte Zustimmungsrate von 60 Prozent in den entsprechenden Antwortkategorien. Der Einsatz der erhobenen alternativen Bewertungsformen wurde häufiger als *eher selten* oder *ganz selten/gar nicht* klassifiziert. Das Lernportfolio wird besonders selten eingesetzt (kumulierte Zustimmungsrate von 60 Prozent in den Kategorien *eher sel-*

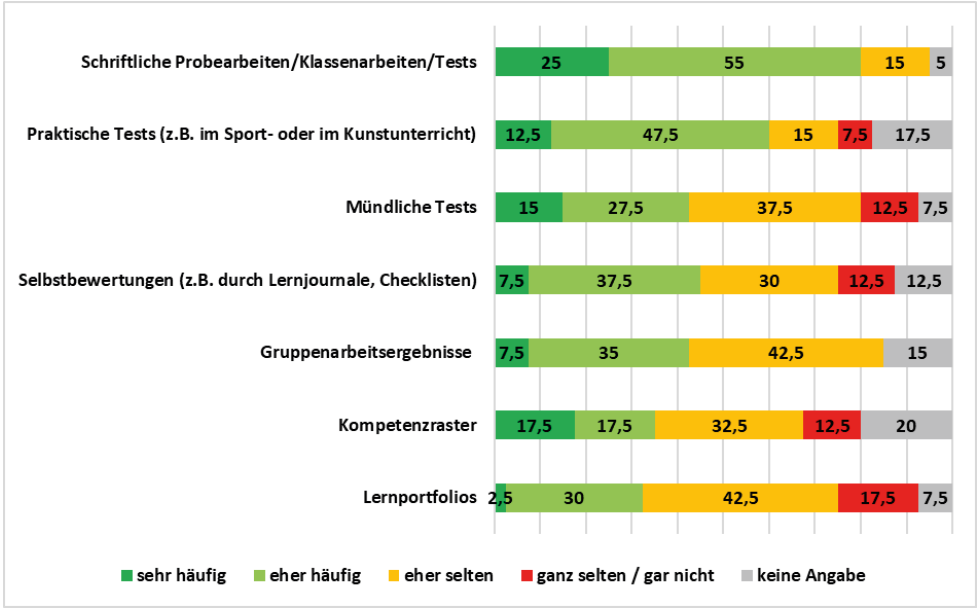

Abbildung 6
Angegebene Einsatzhäufigkeit unterschiedlicher traditioneller und alternativer Bewertungsformen in Prozent (N = 40)

ten oder *ganz selten/gar nicht*). Abbildung 6 gibt die berichtete Einsatzhäufigkeit unterschiedlicher traditioneller und alternativer Bewertungsformen an.

Zusammenhänge zwischen Eignungseinstufung und Einsatzhäufigkeit alternativer Bewertungsformen

Die Korrelationsanalysen zeigten weiterhin, dass Lehrkräfte, die Lernportfolios, Schülerinnen- und Schülerselbstbewertungen und Gruppenarbeitsergebnisse als geeignete Formen der Leistungsbeurteilung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf ansahen, auch statistisch signifikant häufiger von einem Einsatz der jeweiligen Form im Unterricht berichteten. Die entsprechenden Korrelationen können als moderat bis stark klassifiziert werden. Die Zusammenhänge zwischen Eignungseinstufung und Einsatzhäufigkeit wurden über die Rang-Korrelations-Koeffizienten nach Spearman und Kendall quantifiziert und betragen für das Lernportfolio $r_s = .56$ ($p = .002$) bzw. $r_\tau = .53$ ($p = .003$), für die Selbstbewertung $r_s = .48$ ($p = .02$) bzw. $r_\tau = .43$ ($p = .003$) und für Gruppenarbeitsergebnisse $r_s = .57$ ($p = .004$) bzw. $r_\tau = .54$ ($p = .004$). Der Zusammenhang zwischen Eignungseinstufung und Einsatzhäufigkeit des Kompetenzrasters erwies sich hingegen nicht als signifikant ($r_s = .28$, $p = .185$ bzw. $r_\tau = .24$, $p = .188$).

Diskussion

Der vorliegende Beitrag eruierte Potenziale und Stellenwert alternativer Formen der Leistungsbewertung in inklusiven Lernkontexten. Hierbei wurden zunächst gesellschaftliche und pädagogische Funktionen der schulischen Leistungsbewertung differenziert betrachtet (Lintorf, 2012). Im Kontext inklusionsdidaktischer Forderungen treten die entsprechenden Spannungsfelder und Antinomien schulischer Leistungsbewertung verschärft zu Tage (Kopmann, 2016). In der Literatur wird wiederholt auf alternative Bewertungsformen verwiesen, welche zur Verdeutlichung und Anerkennung individueller Lernfortschritte sowie zum selbstregulierten Lernen beitragen sollen (van Bargen, 2017; Streese et al., 2017). Die pädagogische Funktion der Leistungsbewertung zur Optimierung des Lehr- und Lernprozesses sei in den Vordergrund zu rücken (Lintorf, 2012; Walm et al., 2017). Die leistungsförderlichen Effekte ausgewählter alternativer Bewertungsformen stützen sich auf empirische Forschungsbefunde, z.B. in Hinblick auf Schülerinnen- und Schülerselbstbewertungen (Andrade, 2019; Yan et al., 2021), Peer-Bewertungen (Double et al., 2020), Lernportfolios (Chang et al., 2018; Händel et al., 2018) oder das formative Assessment (Schmidt, 2020; Schütze et al., 2018). Der Einsatz von Kompetenzrastern, in denen Kriterien der Leistungsbewertung konkret definiert werden, wirkt sich hierbei in der Regel positiv auf die Reliabilität und Validität von Leistungsbeurteilungen aus (Panadero & Romero, 2014). Ob sich lernförderliche Effekte in konkreten Einzelstudien zeigen bzw. wie hoch diese ausfallen, ist von der Implementierung und dem methodischen Design abhängig.

Vor dem Hintergrund des skizzierten theoretischen Diskurses und empirischen Forschungsstandes untersuchte die vorliegende Studie die Bekanntheit, Eignungseinstufungen und Einsatzhäufigkeit alternativer Bewertungsformen unter 40 Grundschullehrkräften. Die Forschungsfragen können wie folgt beantwortet werden:

1. Angesichts einer seit Jahrzehnten geführten Debatte um die Angemessenheit der Ziffernote wurde von einem hohen prinzipiellen Bekanntheitsgrad alternativer Bewertungsformen ausgegangen. Diese Annahme fand sich bestätigt, wobei alternative Formen der Leistungsbewertung (konkret erhoben wurden: Lernportfolio, Kompetenzraster, Schülerinnen- und Schülerselbstbewertungen, Gruppenarbeitsergebnisse) einer Mehrheit von achtzig bis neunzig Prozent der befragten Lehrpersonen bekannt waren.

2. In der pädagogischen Literatur werden Vorteile alternativer Bewertungsformen propagiert, wie z.B. die angenommene Steigerung von Motivation und Selbstregulation (Bohl, 2019; Grunder & Bohl, 2004). Daher wurde von einer insgesamt recht positiven Eignungseinschätzung alternativer Bewertungsformen bei der vorliegenden pädagogisch vorgebildeten Stichprobe ausgegangen. Die entsprechende Annahme fand sich bestätigt. Das Lernportfolio und das Kompetenzraster wurden in Bezug auf ihre Eignung etwas positiver beurteilt als Selbstbewertungen und die Bewertung von Gruppenarbeitsergebnissen. Die Forschungsliteratur verweist in diesem Kontext darauf, dass insbesondere die Selbsteinschätzungen jüngerer Lernender teilweise weniger valide und daher kritischer zu beurteilen sind (Topping, 2003).
3. Die Eignungseinstufungen unterschiedlicher alternativer Bewertungsformen in Hinblick auf die Leistungsbeurteilung von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf fielen in Bezug auf das Kompetenzraster, Selbstbewertungen und Gruppenarbeitsergebnisse skeptischer aus, wobei maximal die Hälfte der Befragten die entsprechenden Bewertungsformen als sehr geeignet oder eher geeignet beurteilte. Auf deskriptiver Ebene wurde das Lernportfolio zur Leistungsbewertung von Lernenden mit sonderpädagogischem Förderbedarf als vergleichsweise gut geeignet eingestuft. Bei der interferenzstatistischen Prüfung der mittleren Eignungseinstufungen in Bezug auf alle Lernenden im Vergleich zu Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf wurden lediglich Gruppenarbeitsergebnisse als signifikant ungeeigneter in Hinblick auf Lernende mit sonderpädagogischem Förderbedarf bewertet. Die Mittelwertsunterschiede in Hinblick auf die Eignung der weiteren erhobenen alternativen Bewertungsformen unterschieden sich nicht signifikant. Tendenzen, alternative Bewertungsformen aufgrund ggf. erhöhter sprachlicher und kognitiver Anforderungen als eher ungeeignet für Lernende mit besonderem Unterstützungsbedarf zu erachten (vgl. Arnold et al., 2000; Brater et al., 2010), fanden sich dementsprechend nicht bestätigt. Eine Ausnahme stellen Gruppenarbeitsergebnisse dar, deren Eignung insbesondere für Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf kritischer eingestuft wurde. Vermutlich spiegelt sich hierin das Bewusstsein wider, dass Gruppenarbeitsergebnisse von vielen, nicht primär auf die Individualleistung bezogenen Faktoren beeinflusst werden. Beispielswiese können die Dominanz einzelner Gruppenmitglieder, der Trittbrettfahrer-Effekt oder soziale Konflikte in der Gruppe eine Rolle spielen. In Hinblick auf die Eignung alternativer Bewertungsformen für Lernende mit sonderpädagogischem Förderbedarf ist ferner zu berücksichtigen, dass 28 bis 40 Prozent der Befragten keine dezidierte Einstufung vornahmen. Dieser hohe Anteil an Personen ohne entsprechende Angabe könnte auf eine geringe praktische Erfahrung mit den entsprechenden Bewertungsformen hinweisen. Auch die vergleichsweise seltenere Einsatzhäufigkeit alternativer Bewertungsformen würde diese Interpretation nahelegen.
4. Wie erwartet zeigte sich, dass alternative Formen der Leistungsbewertung gegenüber traditionellen Formen (insbesondere schriftlichen Probearbeiten, Klassenarbeiten, Tests) deutlich seltener eingesetzt wurden. Lehrpersonen, die Lernportfolios, Kompetenzraster und Selbstbewertungen als geeignet einstuften, gaben zudem an, die entsprechende Beurteilungsform häufiger einzusetzen. Es fällt ferner auf, dass sowohl das Lernportfolio als auch das Kompetenzraster von vielen Befragten als prinzipiell geeignet eingeschätzt werden.

Dennoch kommen die entsprechenden Beurteilungsinstrumente vergleichsweise selten zum Einsatz. Es ist zu vermuten, dass diesbezüglich insbesondere ein zunächst erhöhter Arbeitsaufwand zur Etablierung alternativer Bewertungsformen eine Rolle spielt (Grittner, 2009; Hebbecker et al., 2020).

Limitationen der Studie

Limitationen der Studie bestehen zum einen in der relativ kleinen, regional begrenzten und vorselektierten Stichprobe, die sich auf Grundschullehrkräfte in einer Großstadt in Nordrhein-Westfalen bezieht und zum anderen darin, dass es sich um Selbstberichte der Lehrkräfte handelt. Einschränkend muss auch genannt werden, dass bei der Befragung nicht nach verschiedenen Förderbedarfen unterschieden wurde. In Bezug auf die Einsatzhäufigkeit der verschiedenen alternativen Bewertungsformen erscheinen die angegebenen Werte selbst bei den beiden „Schlusslichtern" (Kompetenzraster und Lernportfolio) noch relativ hoch, so dass an dieser Stelle auch der Verdacht sozial erwünschter Antworten nicht ausgeschlossen werden kann.

Implikationen für weitere Forschungsprojekte

Weitere Erhebungen zu Eignungseinstufungen und Einsatzhäufigkeit alternativer Bewertungsformen an bundesweiten größeren Stichproben und an unterschiedlichen Schulformen wären wünschenswert. Die Befunde legen insgesamt die Schlussfolgerung nahe, dass die befragten Lehrkräfte trotz weitgehender Kenntnis diverser alternativer Formen der Leistungsbeurteilung doch zu einem großen Teil traditionellere Formen bevorzugen. Ursachen dafür könnten sein, dass die Lehrkräfte eine mangelnde Akzeptanz vonseiten der Schülerinnen und Schüler und Eltern befürchten (Dzelili, 2009; Jachmann, 2003). Diese Vermutung wäre im Rahmen weiterer Forschungsarbeiten zu thematisieren.

Ferner böte sich eine eingehendere Eruierung des Portfolioansatzes an. Das Lernportfolio wurde von den Befragten sowohl im Kontext der allgemeinen Leistungsbewertung als auch in Hinblick auf Lernende mit sonderpädagogischem Unterstützungsbedarf als vergleichsweise geeignet eingestuft. Angesichts der Vielzahl unterschiedlicher Portfolioformen (Häcker, 2011; Schmidt, 2020) wäre in qualitativen Forschungsdesigns weiterführend zu spezifizieren, wie genau Lehrpersonen den entsprechenden Portfolioansatz durchführen und an inklusive Lerngruppen adaptieren.

Implikationen für die Aus- und Fortbildung von Lehrkräften

Die Studienbefunde bieten erste Anregungen zur Gestaltung praktischer Anwendungsfelder. Beispielsweise sollte der zurückhaltende Einsatz alternativer Formen der Leistungsbewertung durch die Lehrkräfte trotz deren Kenntnis im Bereich der Lehrkräfteaus- und -fortbildung aufgegriffen werden, um entsprechende Hindernisse zu identifizieren und diesen gezielt entgegenzuwirken. Eine vermehrte Beschäftigung mit unterschiedlichen Formen der Leistungsbewertung erscheint bereits in der ersten Phase der Lehrkraftausbildung lohnenswert, um Lehramtsstudierende zu einer vertieften Reflexion im Hinblick auf den Einsatz verschiedener Bewertungsformen – auch in Abwägung der Anforderungen eines heterogenitätssensiblen Unterrichts – anzuregen. Die zu konstatierende Diskrepanz zwischen der prinzipiell hohen Eignungseinschätzung und der vergleichsweise seltenen Einsatzhäufigkeit alternativer Bewertungsformen (v.a. Lernportfolio und Kompetenzraster) lässt ferner den Schluss zu, dass kontextspezifisch praktikable Modelle alternativer Beurteilungsformen z.B. im Rahmen der Schul- und Unterrichtsentwicklung anzuregen wären (Sasse & Schulzeck, 2021).

Implikationen für eine inklusive Schulpraxis

In der inklusiven Unterrichtspraxis bieten alternative Formen der Leistungsbewertung Raum, um individuelle Lernprozesse unter Berücksichtigung einer individuellen oder individuell angepassten sachlichen Bezugsnorm zu dokumentieren und zu würdigen (Fischbach et al., 2021; Walm et al., 2017). In Verbindung mit Aspekten des formativen Assessments bzw. einer prozessorientierten Lernverlaufsdiagnostik liefern alternative Bewertungsformen insbesondere auch für Schülerinnen und Schüler mit sonderpädagogischem Unterstützungsbedarf wertvolle Ansatzpunkte zur pädagogischen Förderung (Lenhard, 2014; Mitchell, 2015; Reichenbach & Tiemann, 2018). Inklusionsdidaktische Konzepte binden die schulische Leistungsbewertung in ein formatives pädagogisches Gesamtkonzept ein (Buholzer et al., 2014; Schmidt & Liebers, 2017; Textor, 2015). Alternative Bewertungsformen dienen hierbei als didaktisch-methodische Instrumente oder *teaching tools* im Rahmen eines konstruktivistischen und auf Selbstregulation abzielenden Lernsettings (Anderson, 1998). Vor diesem Hintergrund wäre weiterführend zu eruieren, inwieweit Lehrkräfte alternative Formen der Leistungsbewertung in ihrer Funktion eher als pädagogische Fördermaßnahmen begreifen, wohingegen sie zur summativen Notenvergabe eher traditionelle Beurteilungsformen präferieren. Die Eignung unterschiedlicher Beurteilungsformen ist hierbei immer abhängig von unterschiedlichen pädagogischen oder gesellschaftlichen, formativen oder summativen Funktionen (Lintorf, 2012). Maclellan (2004) konstatiert diesbezüglich, dass alternative Beurteilungsformen in Hinblick auf eine abschließende summative Leistungsbeurteilung und Zertifizierung ebenso problematisch wie tradierte Formen der Leistungsbewertung sind (z.B. mit Bezug auf Objektivitätsmängel). Grunder und Bohl (2004) erklären weiterhin, dass alternative Bewertungsformen in ihrer inneren Struktur oftmals noch komplexer seien als traditionelle.

Die empirisch belegten leistungsförderlichen Effekte formativ orientierter, alternativer Bewertungsformen legen ihren Einsatz auch für den inklusiven Unterricht nahe (Andrade, 2019; Double et al., 2020; Schütze et al., 2018; Yan et al., 2021). Alternative Bewertungsformen unter Partizipation der Lernenden sind hierbei jedoch kein Selbstläufer. Schülerinnen und Schüler benötigen u.a. externes Feedback (Yan et al., 2021) und Training bzw. Mentoring (Bol et al., 2012; Driessen, 2008), um entsprechende alternative Bewertungsformen konstruktiv für ihr weiteres Lernen zu nutzen. Angesichts stark variierender Lernvoraussetzungen in inklusiven Lerngruppen ist ferner eine niveaudifferenzierte Anpassung alternativer Bewertungsformen zu beachten, beispielsweise in Bezug auf kognitive und sprachliche Anforderungen (Dharma & Hermanto, 2019; Sasse & Schulzeck, 2021).

Literatur

Ainscow, M., Booth, T. & Dyson, A. (2006). *Improving schools, developing inclusion*. Routledge.

Allemann-Ghionda, C., Auernheimer, G., Grabbe, H. & Krämer, A. (2006). Beobachtung und Beurteilung in soziokulturell und sprachlich heterogenen Klassen. Die Kompetenzen der Lehrpersonen. *Zeitschrift für Pädagogik, Beiheft 51,* 250–266.

Amrhein, B. & Reich, K. (2014). Inklusive Fachdidaktik. In B. Amrhein & M. Dziak-Mahler (Hrsg.), *Fachdidaktik inklusiv. Auf der Suche nach didaktischen Leitlinien für den Umgang mit Vielfalt in der Schule* (S. 31–44). Waxmann.

Anderson, R. S. (1998). Why Talk About Different Ways to Grade? The Shift from Traditional Assessment to Alternative Assessment. *New Directions for Teaching and Learning, 74*, 5–16.

Andrade, H. L. (2019). A Critical Review of Research on Student Self-Assessment. *Frontiers in Education, 27*, 1–13. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159–181.

Andrade, H. & Valtcheva, A. (2009). Promoting Learning and Achievement Through Self-Assessment. *Theory Into Practice, 48*(1), 12–19.

Arndt, A.-K. & Werning, R. (2017). Leistung an inklusiven Schulen – Perspektiven von Eltern. In A. Textor, S. Grüter, I. Schiermeyer-Reichl & B. Streese (Hrsg.), *Leistung inklusive? Inklusion in der Leistungsgesellschaft. 2. Unterricht, Leistungsbewertung und Schulentwicklung* (S. 130–138). Julius Klinkhardt.

Arnold, K.-H., Froberg, A., Schröder-Begoihn, Ä., Schubert, S. & Vogel, W. (2000). *Integrierte Leistungsbeurteilung in der Orientierungsstufe und Sekundarstufe I. Abschlussbericht Schulbegleitforschungsprojekt 87*. Senator für Bildung und Wissenschaft.

Baume, D. & Yorke, M. (2002). The Reliability of Assessment by Portfolio on a Course to Develop and Accredit Teachers in Higher Education. *Studies in Higher Education, 27*(1), 7–25.

Beutel, S.-I. (2012). Endlich die Noten abschaffen? Ein Plädoyer für die Pädagogisierung der Leistungsbeurteilung. In C. Fischer (Hrsg.), *Diagnose und Förderung statt Notengebung? Problemfelder schulischer Leistungsbeurteilung* (S. 93–106). Waxmann.

Beutel, S.-I. & Pant, H. A. (2019) *Lernen ohne Noten – Alternative Konzepte der Leistungsbeurteilung*. Kohlhammer.

Biglari, A., Izadpanah, S. & Namaziandost, E. (2021). The Effect of Portfolio Assessment on Iranian EFL Learners' Autonomy and Writing Skills. *Education Research International, 2021*. https://doi.org/10.1155/2021/4106882

Binh, N. M. (2021). Portfolio assessment as a tool for promoting reflection in teacher education: a literatur review. *VNU Journal of Science: Foreign Studies*, 37(4), 26–37. https://js.vnu.edu.vn/FS/article/view/4751

Birgin, O. & Baki, A. (2009). An investigation of primary school teachers' proficiency perceptions about measurement and assessment methods: the case of Turkey. *Procedia Social and Behavioral Sciences, 1*, 681–685.

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education, 5*(1), 7–74.

Bohl, T. (2019). Leistungsbewertung, Notengebung und Alternativen zur Notengebung. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Hrsg.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (S. 414–425). Julius Klinkhardt.

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*(4), 280–287. https://doi.org/10.1016/j.cedpsych.2012.02.004

Brater, M., Haselbach, D. & Stefer, A. (2010*). Kompetenzen sichtbar machen. Zum Einsatz von Kompetenzportfolios in Waldorfschulen*. Peter Lang.

Bräu, K. (2018). Inklusion und Leistung. In T. Sturm & M. Wagner-Willi (Hrsg.), *Handbuch schulische Inklusion* (S. 207–221). Babara Budrich.

Brown, G., Andrade, H. & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, G. T. & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Hrsg.), *Handbook of Research on Classroom Assessment* (S. 367–393). Sage. https://dx.doi.org/10.4135/9781452218649

Brunner, I., Häcker, T., & Winter, F. (2008). *Das Handbuch Portfolioarbeit. Konzepte, Anregungen, Erfahrungen aus Schule und Lehrerbildung*. Klett/Kallmeyer.

Buholzer, A., Joller-Graf, K., Kummer Wyss, A. & Zobrist, B. (2014). *Kompetenzprofil zum Umgang mit heterogenen Lerngruppen*. LIT Verlag.

Bürgermeister, A. (2014). Leistungsbeurteilung im Mathematikunterricht: Bedingungen und Effekte von Beurteilungspraxis und Beurteilungsgenauigkeit. *Empirische Erziehungswissenschaft, 45*. Waxmann.

Butler, Y. G. (2018). Young learners' processes and rationales for responding to self-assessment items: cases for generic can-do and five-point Likert-type formats. In J. Davis, J. M. Norris, M. E. Malone, T. H. McKay & Y.-A. Son (Hrsg.), *Useful assessment and evaluation in language education* (S. 21–39). Georgetown University Press.

Chang, C.-C., Liang, C. & Chen, Y.-H. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? *Computers & Education, 60*, 325–334.

Chang, C.-C., Liang, C., Chou, P.-N. & Liao, Y.-M. (2018). Using e-portfolio for learning goal setting to facilitate self-regulated learning of high school students. *Behaviour and Information Technology, 37*, 1237–1251. https://doi.org/10.1080/0144929X.2018.1496275

Chang, C.-C., Tseng, K.-H. & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education, 58,* 303–320.

Chang, C.-C. & Wu, B.-H. (2012). Is Teacher Assessment Reliable or Valid for High School Students under a Web-Based Portfolio Environment? *Educational Technology & Society, 15*(4), 265–278. https://eric.ed.gov/?id=EJ992962

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Buettner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S. & Hardy, I. (2015). Embedded Formative Assessment and Classroom Process Quality: How Do They Interact in Promoting Science Understanding? *American Educational Research Journal, 52*(6), 1133–1159.

De Grez, L., Valcke, M. & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education 13*(2), 129–142. https://doi.org/10.1177/1469787412441284

Dharma, D. S. A. & Hermanto, H. (2019). The implementation of self-assessment of student with special educational needs at inclusive school. *Advances in Social Science, Education and Humanities Research, 296*, 158–162. http://dx.doi.org/10.2991/icsie-18.2019.29

Dietrich, F. (2017). Schulische Inklusion diesseits und jenseits des Leistungsprinzips – Schul- und unterrichtstheoretische Perspektivierungen des Verhältnisses von Inklusion und schulischer Leistungsbewertung. In A. Textor, S. Grüter, I. Schiermeyer-Reichl & B. Streese (Hrsg.), *Leistung inklusive? Inklusion in der Leistungsgesellschaft. 2. Unterricht, Leistungsbewertung und Schulentwicklung* (S. 191–198). Julius Klinkhardt.

Dochy, F., Segers, M. & Sluijsmans, D. (1999). The Use of Self-, Peer and Co-assessment in Higher Education: a review. *Studies in Higher Education, 24*(3), 331–350.

Dori, Y. J. (2003). From Nationwide Standardized Testing to School-Based Alternative Embedded Assessment in Israel: Students' Performance in the Matriculation 2000 Project. *Journal of Research in Science Teaching, 40*(1), 34–52.

Double, K. S., McGrane, J. A. & Hopfenbeck, T. N. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review, 32*, 481–509. https://doi.org/10.1007/s10648-019-09510-3

Driessen, E. W. (2008). *Educating the self-critical doctor: using a portfolio to stimulate and assess medical students reflection*. Universiteit Maastricht. https://doi.org/10.26481/dis.20080625ed

Driessen E. W., Overeem, K., van Tartwijk, J., van der Vleuten C. P. M. & Muijtjens, A M. M. (2006). Validity of portfolio assessment: which qualities determine ratings? *Medical Education, 40*, 862–866. https://doi.org/10.1111/j.1365-2929.2006.02550.x

Driessen, E. W., van der Vleuten, C. P. M., Schuwirth, L., van Tartwijk, J. & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Medical Education, 39*, 214–220. https://doi.org/10.1111/j.1365-2929.2004.02059.x

Dzelili, A. (2009). Noten gehören verboten – aber warum? *Bildung Schweiz, 1*, 8–11.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. https://doi.org/10.2307/1170785

Falkenberg, K. (2020). *Gerechtigkeitsüberzeugungen bei der Leistungsbeurteilung. Eine Grounded-Theory-Studie mit Lehrkräften im deutsch-schwedischen Vergleich*. Springer VS.

Falkenberg, K., Vogt, B. & Waldow, F. (2017). Ständig geprüft oder kontinuierlich unterstützt? Schulische Leistungsbeurteilung in Schweden zwischen formativem Anspruch und summativer Notwendigkeit. *Zeitschrift für Pädagogik, 63*(3), 317–333.

Favier, R. P., Vernooij, J. C. M., Jonker, F. H. & Bok, H. G. J. (2019). Inter-Rater Reliability of Grading Undergraduate Portfolios in Veterinary Medical Education. *Journal of Veterinary Medical Education, 46*(4), 415–422. https://doi.org/10.3138/jvme.0917-128r1

Fischbach, A., Mähler, C. & Hasselhorn, M. (2021). Grundlagen der Diagnostik im inklusiven Kontext. In C. Mähler & M. Hasselhorn (Hrsg.), *Inklusion. Chancen und Herausforderungen* (S. 85–98). Hogrefe.

Fischer, C. (Hrsg.). (2012*). Diagnose und Förderung statt Notengebung*. Waxmann.

Frohn, J. (2019). Kompetenzorientierung und Inklusion – eine Zusammenführung auf Unterrichtsebene. *Herausforderung Lehrer_innenbildung*, *2*(1), 15–38. https://doi.org/10.4119/UNIBI/hlz-48

Fürstenau, S. & Gomolla, M. (2012). *Migration und schulischer Wandel: Leistungsbeurteilung*. Springer VS.

Gadbury-Amyot, C. C., McCracken, M. S., Woldt, J. L. & Brennan, R. L. (2014). Validity and reliability of portfolio assessment of student competence in two dental school populations: a four-year study. *Journal of Dental Education*, *78*(5), 657–667. https://doi.org/10.1002/j.0022-0337.2014.78.5.tb05718.x

Gikandi, J. W., Morrow, D. & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*, 2333–2351.

Gold, A. (2011). *Lernschwierigkeiten. Ursachen, Diagnostik, Intervention*. Kohlhammer.

Graham, S., Hebert, M. & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal, 115*(4), 523–547.

Grittner, F. (2009). *Leistungsbewertung mit Portfolio in der Grundschule. Eine mehrperspektivische Fallstudie aus einer notenfreien sechsjährigen Grundschule*. Julius Klinkhardt. https://doi.org/10.25656/01:2026

Grunder, H.-U. & Bohl, T. (Hrsg.). (2004). *Neue Formen der Leistungsbeurteilung*. Schneider Hohengehren.

Häcker, T. (2011). Portfolioarbeit – Ein Konzept zur Wiedergewinnung der Leistungsbeurteilung für die pädagogische Aufgabe der Schule. In W. Sacher & F. Winter (Hrsg.), *Diagnose und Beurteilung von Schülerleistungen. Grundlagen und Reformansätze* (S. 217–230). Schneider Hohengehren.

Händel, M., Wimmer, B. & Ziegler, A. (2018). E-portfolio use and its effects on exam performance – a field study. *Studies in Higher Education, 45*(2), 258–270. https://doi.org/10.1080/03075079.2018.1510388

Harks, B., Rakoczy, K., Hattie, J., Besser, M. & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology: An International Journal of Experimental Educational Psychology, 34*(3), 269–290.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112.

Hebbecker, K., Förster, N., Forthmann, B., Heyne, L., Peters, M. T., Salaschek, M. et al. (2020). Diagnostik, Feedback und differenzierte Leseförderung. Umsetzung evidenzbasierter Konzepte im schulischen Alltag. *Leseforum.ch*, *3*. https://www.leseforum.ch/sysModules/obxLeseforum/Artikel/710/2020_3_de_hebbecker_et_al.pdf

Heller, J. I., Sheingold, K. & Myford, C. M. (1998). Reasoning about evidence in portfolios: cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*, 5–40.

Herman, J. I., Gearhart, M. & Baker, E. I. (1993). Assessing writing portfolios: issues in the validity and meaning of scores. *Educational Assessment, 1*, 201–224.

Heydrich, J., Weinert, S., Nusser, L., Artelt, C. & Carstensen, C. H. (2013). Die Einbeziehung von Förderschülern in Large-Scale-Kompetenzerhebungen: Herausforderungen und Vorgehen im Rahmen des Nationalen Bildungspanels (NEPS). *Journal for educational research online, 5*(2), 217–240. https://doi.org/10.25656/01:8431

Holder, K. & Kessels, U. (2019). Unterrichtsgestaltung und Leistungsbeurteilung im inklusiven und standardorientierten Unterricht aus der Sicht von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 22*, 325–346.

Hollenbach-Biele, N. & Klemm, K. (2020). *Inklusive Bildung zwischen Licht und Schatten: Eine Bilanz nach zehn Jahren inklusiven Unterrichts*. Bertelsmann Stiftung.

Isnawati, I., Yulianti, D. & Samhati, S. (2021). Portfolio assessment as a problem based learning model to help elementary school students' deal with Mathematics. *International Journal of Educational Studies in Social Sciences, 1*(3), 110–113. https://doi.org/10.53402/ijesss.v1i3.25

Jachmann, M. (2003). *Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern*. Leske + Budrich.

Joller-Graf, K. (2010). Binnendifferenzierter Unterricht. In A. Buholzer & A. Kummer Wyss (Hrsg.), *Alle gleich – alle unterschiedlich. Zum Umgang mit Heterogenität in Schule und Unterricht* (S. 122–137). Klett/Kallmeyer.

Kaul, M. (2020). *Tiergestützte Interventionen als Beitrag zur Umsetzung von Inklusion in der Schule. Eine Untersuchung der Heilpädagogischen Förderung mit dem Pferd an Grundschulen in Münster*. Dissertation, Universität Münster. https://nbn-resolving.org/urn:nbn:de:hbz:6-00189612906

Kingston, N. & Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Klemm, K. (2018). *Unterwegs zur inklusiven Schule Lagebericht 2018 aus bildungsstatistischer Perspektive*. Bertelsmann Stiftung.

KMK (Kultusministerkonferenz) (2022). *Sonderpädagogische Förderung in Schulen 2011 bis 2020*. Sekretariat der Ständigen Konferenz der Kultusminister. https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Dok231_SoPaeFoe_2020.pdf

Kopmann, H. (2016). *Einstellungen und Kompetenzen von Lehrpersonen im Kontext inklusiver Bildung: Eine empirische Erhebung inklusionsrelevanter Einstellungskonstrukte auf Lehrkraftseite und des schülerperzipierten Klassenklimas*. Veröffentlichte Dissertation, Westfälische Wilhelms-Universität Münster. https://miami.uni-muenster.de/Record/0cce0136-b078-4c21-989d-0fe5ad3781d8

Kutzer, R. (1982). Anmerkungen zum Struktur- und Niveauorientierten Unterricht. In H. Probst (Hrsg.), *Kritische Behindertenpädagogik in Theorie und Praxis* (S. 29–62). Jarik.

Lau, R., & Lübeck, A. (2021). Notengebung auf Wunsch? Zieldifferente Leistungsbewertung im Spannungsfeld von (vermuteten) Bedürfnissen und realen Konsequenzen. *DiMawe – Die Materialwerkstatt, 3*(2), 38–48. https://doi.org/10.11576/dimawe-4127

Lawrenz, F., Huffman, D. & Welch, W. (2001). The Science Achievement of Various Subgroups on Alternative Assessment Formats. *Science Education, 85*(3), 279–290.

Lehmann, R. H., Peek, R. & Gänsfuß, R. (1997). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen*. Hamburg: Behörde für Schule, Jugend und Berufsbildung.

Lehmann, R. H., Peek, R. & Gänsfuß, R. (1999). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern an Hamburger Schulen – Klassenstufe 7*. Behörde für Schule, Jugend und Berufsbildung.

Lenhard (2014). Leistungsmessung und Leistungsbewertung. In F. B. Wember, R. Stein, U. Heimlich (Hrsg.), *Handlexikon Lernschwierigkeiten und Verhaltensstörungen* (S. 148–150). Kohlhammer.

Li, H., Xiong, Y., Hunter, C. V., Guo, X. & Tywoniw, R. (2020). Does peer assessment promote student learning? *A meta-analysis*. *Assessment & Evaluation in Higher Education*, *45*(2), 193–211. https://doi.org/10.1080/02602938.2019.1620679

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S. et al. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746

Li, M. & Zhang, X. (2020). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Springer VS. https://doi.org/10.1007/978-3-531-94339-8

Liu, X. & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education*, *39*(3), 275–292. https://doi.org/10.1080/02602938.2013.823540

Maclellan, E. (2004). How convincing is alternative assessment for use in higher education? *Assessment & Evaluation in Higher Education, 29*(3), 311–321.

Maier, U. (2015). *Leistungsdiagnostik in Schule und Unterricht. Schülerleistungen messen, bewerten und fördern*. Julius Klinkhardt.

Maier, U. (2019). Formative Leistungsdiagnostik. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Hrsg.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (S. 403–413). Julius Klinkhardt.

Mak, P. & Wong, K. M. (2018). Self-regulation through portfolio assessment in writing classrooms. *ELT Journal*, *72*(1), 49–61. https://doi.org/10.1093/elt/ccx012

Maslovaty, N. & Kuzi, E. (2002). Promoting motivational goals through alternative or traditional assessment. *Studies in Educational Evaluation, 28*, 199–222.

McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning*, *33*(6), 562–574. https://doi.org/10.1111/jcal.12200

McMillan, J. H., Venable, J. C. & Varier, D. (2013). Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed. *Practical Assessment, Research, and Evaluation, 18*(2). https://doi.org/10.7275/tmwm-7792

Meeus, W., van Petegem, P. & Engels, N. (2009). Validity and reliability of portfolio assessment in pre-service teacher education. *Assessment & Evaluation in Higher Education, 34*(4), 401–413. https://doi.org/10.1080/02602930802062659

Middendorf, W. (2012). Schulische Leistungsbeurteilung auf dem Prüfstand: einführende Betrachtungen zu aktuellen Aufgaben und Herausforderungen. In C. Fischer (Hrsg.), *Diagnose und Förderung statt Notengebung? Problemfelder schulischer Leistungsbeurteilung* (S. 9–22). Waxmann.

Mitchell, D. (2014). *What really works in special and inclusive education: Using evidencebased teaching strategies* (2. Aufl.). Routledge.

Mitchell, D. (2015). *Education that fits: Review of international trends in the education of students with special educational needs*. University Of Canterbury. https://www.education.vic.gov.au/Documents/about/department/psdlitreview_Educationthatfits.pdf

Mitchell, D., Morton, M. & Hornby, G. (2010). *Review of the literature on Individual Education Plans. Report to the New Zealand Ministry of Education*. College of Education, University of Canterbury.

Muho, A. & Leka, K. (2021). Students' Perceptions of Portfolio as a Motivating Factor in Learning English as a Foreign Language. *Journal of Educational and Social Research, 11*(6), 47. https://doi.org/10.36941/jesr-2021-0127

Nagel, M. & Lindsey, B. (2018). The use of classroom clickers to support improved self-assessment in introductory chemistry. *Journal of College Science Teaching*, *47*(5), 72–79.

Neumann, P. & Lütje-Klose, B. (2020). Diagnostik in inklusiven Schulen – zwischen Stigmatisierung, Etikettierungs-Ressourcen-Dilemma und förderorientierter Handlungsplanung. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer & P. Stanat (Hrsg.), *Schüler*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 3–28). Springer VS.

Nezakatgoo, B. (2011). The Effects of Portfolio Assessment on Writing of EFL Students. *English Language Teaching, 4*(2), 231–241.

Nicol, D. J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nystrand, M. Cohen, A. S. & Dowling, N. M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1*, 53–70.

Oldfield, K. A. & Malcapine, J. M. K. (1995). Peer and self-assessment at the tertiary level - an experiential report. *Assessment and Evaluation in Higher Education, 20*, 125–132.

Orsmond, P., Merry, S. & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21*, 239–249.

Panadero, E. & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E. & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 133–148. https://doi.org/10.1080/0969594X.2013.877872

Panadero, E., Romero, M. & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H. & Weinert, S. (2016). Testing Students with Special Educational Needs in Large-Scale Assessments – Psychometric Properties of Test Scores and Associations with Test Taking Behavior. *Frontiers in Psychology, 7*, 1–14. https://doi.org/10.3389/fpsyg.2016.00154

Prengel, A. (2006). *Pädagogik der Vielfalt. Verschiedenheit und Gleichberechtigung in Interkultureller, Feministischer und Integrativer Pädagogik* (3. Aufl.). VS Verlag für Sozialwissenschaften.

Prengel, A. (2007). Kinder im Prisma der Lehrkraftwahrnehmung – Verfahren der Leistungs- und Entwicklungsdokumentation an der Montessori-Gesamtschule Potsdam. In J. Hofmann (Hrsg.), *Neue Formen des Lehrens und Lernens. Leistungsbewertung ohne Zensuren und jahrgangsübergreifender Unterricht in der Montessori-Gesamtschule Potsdam* (S. 20–55). Julius Klinkhardt.

Rakoczy, K. (2012). Formatives Assessment – theoretische Erkenntnisse und praktische Umsetzung im Mathematikunterricht. In C. Fischer (Hrsg.), *Diagnose und Förderung statt Notengebung? Problemfelder schulischer Leistungsbeurteilung* (S. 73–92). Waxmann.

Rakoczy, K., Klieme, E., Bürgermeister, A. & Harks, B. (2008). The Interplay Between Student Evaluation and Instruction. *Zeitschrift für Psychologie*, *216*(2), 111–124. https://doi.org/10.1027/0044-3409.216.2.111

Reichenbach, C. & Thiemann, H. (2018). *Lehrbuch diagnostischer Grundlagen der Heil- und Sonderpädagogik*. Verlag modernes Lernen.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H. & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066. https://doi.org/10.1037/edu0000190

Sasse, A. & Schulzeck, U. (2014). Von der Schülerleistung zur Leistungsbewertung im Gemeinsamen Unterricht – erneuter Zwischenstand in einem Schulversuch. In Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien (Hrsg.), *Unterricht im Spannungsfeld zwischen Kompetenz- und Standardorientierung*. (S. 38–57). SDC Satz + Druck.

Sasse, A. & Schulzeck, U. (2021). *Inklusiven Unterricht planen, gestalten und reflektieren. Die Differenzierungsmatrix in Theorie und Praxis*. Klinkhardt.

Scheunpflug, A., Stadler-Altmann, U. & Zeinz, H. (2012). *Bestärken und fördern - Wege zu einer veränderten Kultur des Lernens in der Sekundarstufe I*. Kallmeyer in Verbindung mit Klett.

Schmidinger, E. (2012). Lern- und unterrichtstheoretische Begründung alternativer Formen der Leistungsbeurteilung. In F. Hellmich, S. Förster & F. Hoya (Hrsg.), *Bedingungen des Lehrens und Lernens in der Grundschule Bilanz und Perspektiven* (S. 101–104). Springer VS.

Schmidinger, E. (2013). Formative Leistungsbeurteilung. *Erziehung & Unterricht*, 163, S. 776–785.

Schmidinger, E., Hofmann, F. & Stern, T. (2016). Leistungsbeurteilung unter Berücksichtigung ihrer formativen Funktion. In M. Bruneforth, F. Eder, K. Krainer, C. Schreiner, A. Seel & C. Spiel (Hrsg.), *Nationaler Bildungsbericht Österreich 2015. Band 2. Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (S. 59–94). Leykam.

Schmidt, C. (2020). *Formatives Assessment in der Grundschule. Konzept, Einschätzungen der Lehrkräfte und Zusammenhänge*. Springer VS.

Schmidt, C. & Liebers, K. (2017). Formatives Assessment im inklusiven Unterricht – Forschungsstand und erste Befunde. In F. Hellmich & E. Blumberg (Hrsg.), *Inklusiver Unterricht in der Grundschule* (S. 50–65). Kohlhammer.

Schmude, C. (2001). *Berichtszeugnis – unnötiger Aufwand oder aufwendige Notwendigkeit?* Dissertation, Humboldt-Universität Berlin.

Schuck, K. D. (2014). Individualisierung und Standardisierung in der inklusiven Schule - ein unauflösbarer Widerspruch? *Die Deutsche Schule, 106*(2), 162–174.

Schütze, B., Souvignier, E. & Hasselhorn, M. (2018). Stichwort - formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*(4), S. 697–715.

Segaran, M., & Hasim, Z. (2021). Self-regulated learning through e-portfolio: A meta-analysis. *Malaysian Journal of Learning and Instruction*, *18*(1), 131–156. https://doi.org/10.32890/mjli2021.18.1.6

Simon, J. & Simon, T. (2014). Inklusive Diagnostik – Wesenszüge und Abgrenzung von traditionellen „Grundkonzepten" diagnostischer Praxis. Eine Diskussionsgrundlage. *Zeitschrift für Inklusion*, (4). https://www.inklusion-online.net/index.php/inklusion-online/article/view/194

Siow, L.-F. (2015). Students' perceptions on self- and peer-assessment in enhancing learning experience. *Malaysian Online Journal of Educational Sciences*, *3*(2), 21–35.

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in Education and Teaching International*, *39*(1), 71–81.

Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz: Eine empirische Studie in der Sekundarstufe I. Schul- und Unterrichtsforschung: Bd. 10*. Schneider Hohengehren.

Stiggins, R. J., Arter, J. A., Chappuis, J. & Chappuis, S. (2007). *Classroom assessment for student learning. Doing it right – using it well* (2nd Edition). Pearson Education.

Streese, B., Schiermeyer-Reichl, I., Meyer, A., Moritz, F. & Wenzel, E. (2017). Inklusiv unterrichten – inklusiv bewerten? Impulse zur ‚inklusiven Leistungsbewertung' in Schulen der Sekundarstufe. In A. Textor, S. Grüter, I. Schiermeyer-Reichl & B. Streese (Hrsg.), *Leistung inklusive? Inklusion in der Leistungsgesellschaft. 2. Unterricht, Leistungsbewertung und Schulentwicklung* (S. 121–129). Julius Klinkhardt.

Sturm, T. (2015). Inklusion: Kritik und Herausforderung des schulischen Leistungsprinzips. *Erziehungswissenschaft, 26*(51), 25–32.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Taras, M. (2010). Assessment for learning: assessing the theory and evidence. *Procedia Social and Behavioral Sciences, 2*, 3015–3022.

Textor, A. (2015). *Einführung in die Inklusionspädagogik*. Klinkhardt.

Tiara Linanti, A., Ridlo, S., & Bintari, S. H. (2021). The Implementation of Portfolio Assessment to Increase Critical Thinking Ability for High School Students on Human Coordination System Material. *Journal of Innovative Science Education*, *10*(2), 130–136. https://doi.org/10.15294/jise.v9i3.41065

Tochel, C., Haig, A., Hesketh, A., Cadzow, A., Beggs, K., Colthart, I. et al. (2009). The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. *Medical Teacher, 31*, 299–318.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Hrsg.), *Optimising new modes of assessment: In search of qualities and standards* (S. 55–87). Kluwer Academic Publishers.

Valtin, R (2012). Noten oder verbale Beurteilungen: Was ist ein gutes Zeugnis? In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Leistungsbeurteilung* (S. 89–106). Springer VS.

Van der Gulden, R., Heeneman, S., Kramer, A. W. M., Laan, R. F. J. M., Scherpbier-de Haan, N. D. & Thoonen, B. P. A. (2020). How is self-regulated learning documented in e-portfolios of trainees? A content analysis. *BMC Medical Education, 20*. https://doi.org/10.1186/s12909-020-02114-4

Van Zundert, M., Sluijsmans, D. & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Von Bargen, I. (2017). Zum Leistungsverständnis von Lehrkräften im inklusiven Alltag – Einblicke in eine qualitative Längsschnittstudie. In A. Textor, S. Grüter, I. Schiermeyer-Reichl & B. Streese (Hrsg.), *Leistung inklusive? Inklusion in der Leistungsgesellschaft. 2. Unterricht, Leistungsbewertung und Schulentwicklung* (S. 148–156). Julius Klinkhardt.

Wagner, C. & Valtin, R. (2003). Noten oder Verbalbeurteilungen? Die Wirkung unterschiedlicher Bewertungsformen auf die schulische Entwicklung von Grundschulkindern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 35*(1), 27–36.

Walm, M., Schultz, C., Häcker, T. & Moser, V. (2017). „Diagnostik und Leistungsbewertung im Dienste des Lernens" – Theoretische Perspektiven auf ein inklusives Entwicklungsfeld. In A. Textor, S. Grüter, I. Schiermeyer-Reichl & B. Streese (Hrsg.), *Leistung inklusive? Inklusion in der Leistungsgesellschaft. 2. Unterricht, Leistungsbewertung und Schulentwicklung* (S. 113–120). Julius Klinkhardt.

Wang, W. (2017). Using rubrics in student self-assessment: student perceptions in the English as a foreign language writing context. *Assessment & Evaluation in Higher Education*, *42*(8), 1280–1292. https://doi.org/10.1080/02602938.2016.1261993

Wanner, T. & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education*, *43*(7), 1032–1047. https://doi.org/10.1080/02602938.2018.1427698

Watt, H. M. G. (2005). Attitudes to the Use of Alternative Assessment Methods in Mathematics: A Study with Secondary Mathematics Teachers in Sydney, Australia. *Educational Studies in Mathematics, 58*, 21–44.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Beltz.

Wild, K.-P. & Krapp, A. (2006). Pädagogisch-psychologische Diagnostik. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 525-574). Beltz/PVU.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3–14.

Winter, F. (2004). *Leistungsbewertung – Eine neue Lernkultur braucht einen anderen Umgang mit den Schülerleistungen*. Schneider Hohengehren.

Winter, F. (2012). Klassenarbeit passé? In C. Fischer (Hrsg.), *Diagnose und Förderung statt Notengebung? Problemfelder schulischer Leistungsbeurteilung* (S. 57–72). Waxmann.

Yan, Z., Wang, X., Boud, D. & Lao, H. (2021). *The effect of self-assessment on academic performance and the role of explicitness*: A meta-analysis. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.2012644

Yilmaz, F. N. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice, 17*, 395–409. https://doi.org/10.12738/estp.2017.2.0098

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M. et al. (2008). On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change. *Applied Measurement in Education*, *21*(4), 335–359. https://doi.org/10.1080/08957340802347845

Zundert, M., Sluijsmans, D. & Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Autorinnen- und Autorenhinweis:

Korrespondenzadresse:
Dr. Henrike Kopmann
Fachbereich 06 Erziehungswissenschaften und Sozialwissenschaften
Westfälische Wilhelms-Universität Münster
Georgskommende 33
D-48143 Münster
henrike.kopmann@uni-muenster.de

Erstmals eingereicht: ??.??.2021
Überarbeitung eingereicht: 11.04.2022
Angenommen: 29.04.2022