Spe Schulpraxis entwickeln
Journal für forschungsbasierte Schulentwicklung

Systematische Literaturrecherche zur Nutzung von standardisierten Leistungstests im Unterricht

Helena Dorothea Zenker

Technische Universität Dresden, Studentin
forus@tu-dresden.de

Zusammenfassung:
Standardisierte Leistungstests sind hauptsächlich im Rahmen von (inter)nationalem Bildungsmonitoring im deutschen Bildungssystem etabliert. Daneben bieten sie auch die Möglichkeit, Lehrpersonen wichtige Rückmeldungen über den Lernfortschritt der Schüler:innen zu geben und somit Einfluss auf die Unterrichtsgestaltung zu nehmen. In der vorliegenden Arbeit wird eine systematische Literaturrecherche zur formativen Nutzung standardisierter Leistungstests im Schul- beziehungsweise Unterrichtskontext aufgezeigt. Zu diesem Zweck wurde eine Datenbankrecherche mithilfe der Datenbanken *PSYNDEXplus*, *FIS Bildung*, *ERIC* und *Academic Search Elite* durchgeführt. Es konnten drei Anwendungsbereiche standardisierter Tests im formativen Kontext, zur *Früherkennung von Schullaufbahngefährdung (1)*, *zur Überprüfung der Effektivität und Vorhersagekraft von Schüler:innenleistung (2)* und *zur Vorbereitung auf summative standardisierte Leistungstests (3)*, kategorisiert werden.
Die Ergebnisse zeigen, dass standardisierte Leistungstests und formative Assessments zunehmend computergestützt und damit automatisiert sowie zeiteffektiv durchgeführt werden, um Lehrkräfte in der Leistungsdiagnostik zu entlasten. Die Ergebnisse implizieren, dass Lehrkräfte professionelle Defizite in der Nutzung und Analyse (digitaler) Test- und Assessment-Daten haben. Basierend auf dem aktuellen Forschungsstand ist der Schwerpunkt unterrichtlicher Leistungsüberprüfung prospektiv eher auf verschiedenen Methoden und Modi formativen Assessments zu verorten als auf der expliziten Nutzung von standardisierten Leistungstests.

Schlagwörter:
Systematische Literaturrecherche, Formatives Assessment, standardisierte Leistungstests

Summary:
Standardized achievement tests are mainly established within the framework of (inter)national educational monitoring in the German education system. In addition, they also offer the opportunity to provide teachers with important feedback on the learning progress of their progress and thus influence the design of lessons. This paper presents a systematic literature review on the formative use of standardized achievement tests in the school or classroom context. For this purpose, a database search was carried out using the PSYNDEXplus, FIS Bildung, ERIC and Academic Search Elite databases. It was possible to categorize three areas of application of standardized tests in a formative context, for the early detection of school career risk (1), for checking the effectiveness and predictive power of student performance (2) and for preparing for summative standardized achievement tests (3). The results show that standardized achievement tests and formative assessments are increasingly computer-based and thus automated and time-effective in order to relieve teachers in performance diagnostics. The results imply that teachers have professional deficits in the use and analysis of (digital) test and assessment data. Based on the current state of research, the focus of classroom performance assessment is prospectively more likely to be on various methods and modes of formative assessment than on the explicit use of standardized achievement tests.

Keywords: systematic literature review, formative assessment, standardized achievement tests

1 Einleitung

Standardisierte Leistungstests bieten die Möglichkeit, den Leistungsstand von Schüler:innen auf individueller Ebene objektiv abzubilden, da es sich um vergleichbare, an einer Stichprobe entwickelte Tests mit einem normierten Ablauf und einer geeichten Auswertung handelt (vgl. Schwerdt, 2010, S. 14 f.). In Form internationaler Schulleistungsstudien, wie PISA-, TIMSS-, und IGLU oder nationaler Vergleichsarbeiten (VERA) sind standardisierte Leistungstests bereits im deutschen Bildungssystem etabliert. Bezogen auf die Unterstützung individueller Entwicklung der Schüler:innen sind die Zielvorstellungen solcher Tests jedoch kritisch zu hinterfragen:

„Standardisierte Leistungsstudien in der Schule können als ein Ausdruck von Anforderungen der Gesellschaft an die Schüler:innen gelten, denn die Definition von Kompetenzniveaus und Konstruktion von Testaufgaben ist Ausdruck dessen, was von jungen Menschen heute erwartet wird. Diese Erwartungen besitzen aufgrund des öffentlichen Vergleichscharakters der Studien hohe Verbindlichkeit“ (Schwerdt, 2010, S. 14).

Grundsätzlich haben standardisierte Leistungsmessungen und (inter)nationale Vergleichsstudien im Kontext von Bildungsmonitoring einen (bildungs)politischen Fokus. Gleichzeitig ermöglichen standardisierte Tests, vergleichbare Daten über Schüler:innenkompetenzen und -kenntnisse zu erlangen. Sie werden zwar häufig unter Mitwirkung der Lehrkräfte durchgeführt, jedoch wurde die Nutzung der Informationen über individuelle Schüler:innenleistungen für die Unterrichtsgestaltung bisher wenig elaboriert (Souvignier et al. 2021). Die vorliegende Arbeit beschäftigt sich mit der Frage nach einer erfolgreichen Nutzbarmachung von standardisierten Leistungstests, nicht nur zur bloßen Notengebung, beziehungsweise als Qualifikationsnachweis, sondern als diagnostisches Instrument zur Nachverfolgung und Evaluation von Lernentwicklungen. Daher soll mithilfe dieses systematischen Reviews der Betrachtungsschwerpunkt auf die Unterrichtsebene, also die Mikroebene, gelegt werden. Ziel der Arbeit ist es demnach herauszufinden, welche aktuellen Forschungsergebnisse zur formativen Anwendung standardisierter Leistungstests in Schule mit Fokus auf den Unterricht existieren, die von Lehrpersonen zur Weiterentwicklung von Lernangeboten genutzt werden können. Weiterführend wird nach Bedingungen gefragt, welche nötig sind, damit standardisierte Leistungstests zur Unter-stützung des Lernens genutzt werden können[1].

Im Beitrag wird zunächst auf den Bereich Leistungsmessung und des Konzepts des formativen Assessments respektive der Lernverlaufsdiagnostik eingegangen und die sich daraus ergebende Fragestellung formuliert. Es folgt die Darstellung der Methodik zur Literaturrecherche (Kap. 3), Ergebnisdarstellung (Kap. 4) und Diskussion (Kap.5).

2 Leistungsbeurteilung in der Schule

Die Leistungsbeurteilung im schulischen Kontext ist ein essenzieller diagnostischer Bestandteil des Unterrichts. Sie dient nicht nur der Beurteilung von Leistungen von Schüler:innen im Rahmen von Bewertungsvorgängen mit Fokus auf den Resultaten eines Lernprozesses, sondern betrachtet ebenso die Rückmeldung über die individuelle Leistungs- und Kompetenzentwicklung beziehungsweise den Lernfortschritt der Schüler:innen. Somit stellt die Leistungsbeurteilung gleichzeitig einen wichtigen Faktor

[1] Im Verlauf der Recherche musste die ursprüngliche Fragestellung (Nutzung standardisierter Leistungstests im Unterricht durch Lehrkräfte) abstrahiert werden, da zur expliziten Nutzung standardisierter Tests im Unterricht Ergebnisse ausblieben. Folglich bezieht sich das Review übergeordnet auf die formative Nutzung standardisierter Tests im allgemeineren Schulkontext.

für die Unterrichtsreflexion sowie -gestaltung dar, um heterogenen Ansprüchen an die unterrichtliche Didaktik und Methodik gerecht zu werden (vgl. Jürgens, 2022, S. 552).

„Die Ausübung der Leistungsbeurteilung erfolgt somit als eine professionelle Tätigkeit, die einerseits als förderorientierte Bewertungspraxis Feedback zu individuellen Lern- und Leistungsentwicklungen gibt, andererseits als eignungsbezogene Evaluationsangelegenheit Laufbahnentscheidungen von Schülerinnen und Schülern vorbereitet und ausführt sowie Abschlusszertifikate vergibt“ (Jürgens, 2022, S. 552).

2.1 Funktionen der Leistungsbeurteilung bzw. -diagnostik

Die Leistungsbeurteilung erfüllt zwei Hauptfunktionen. Im gesellschaftlichen Sinne dient sie der Selektion, da die Schule zentral in diversen Platzierungs- und Eignungsverfahren, beispielsweise Schulleistungstests oder Schulabschlusstests, fungiert. Daher bestimmen Leistungsbeurteilungen in Schulen unter anderem weitere Bildungslaufbahnen und -chancen. Jedoch erfüllt die Schule ebenso einen Entwicklungsanspruch, wobei Leistungsbeurteilungen zur Diagnostik individueller Entfaltungsmöglichkeiten beziehungsweise Lernschwächen genutzt werden. (vgl. Jürgens, 2022, S. 554 f.)

Je nachdem, ob die Beurteilung eine gesellschaftliche oder didaktische Funktion erfüllen soll, werden unterschiedliche Verfahrensweisen notwendig. Man unterscheidet zwischen formativer und summativer Diagnostik. Während formative Diagnostik Leistungen der Schüler:innen im Lern*prozess* betrachtet, stehen bei der summativen Diagnostik Lern-*ergebnisse* im Fokus. Deswegen wird letztere insbesondere für die Feststellung von Qualifikationen genutzt. Denn für die Entscheidung und Zuordnung von Individuen zu weiteren Bildungsangeboten und Berufen ist letztendlich nur die Beurteilung des Endproduktes relevant. Summative Leistungsdiagnosen erfolgen demzufolge beispielsweise nach Abschluss komplexer Unterrichtsreihen und -einheiten oder am Ende einer Schullaufbahn oder eines Schuljahres, also in großen Zeitabständen. Die formative Diagnostik dagegen greift gestaltend in den Lernprozess ein, indem die Leistungsbeurteilungen hierbei in kürzeren Abständen erfolgen, um innerhalb von Unterrichtseinheiten Lernbarrieren sowie Schwierigkeiten in der Wissensvermittlung aufzudecken und direkt mit passenden Förderstrategien und Unterrichtsadaptionen darauf einzugehen. Deshalb spricht man hierbei von Lernprozessdiagnostik. Die kontinuierliche unterrichtsbegleitende Kontrolle der zu erreichenden fachbezogenen und überfachlichen Kompetenz- und Lernziele offenbart außerdem, inwieweit erfolgte pädagogisch-didaktische Anpassungen erfolgreich waren oder sind. (vgl. Jürgens, 2022, S. 556 f.).

2.2 Standardisierte Leistungstests im Rahmen der Kompetenzdiagnostik

Nachdem bis zum Ende der 1990er Jahre der Interessenschwerpunkt auf der Wissensvermittlung im Sinne von Input lag, orientiert an den curricularen Vorgaben und Schulgesetzen, gewann mit der Jahrtausendwende, durch die Etablierung der Schulpsychologie, die Intelligenz- und Leistungsdiagnostik an Priorität. Folglich rückten vielmehr die Ergebnisse, also der Output, von Bildungsprozessen in Form von messbaren Kompetenzen in den Vordergrund (vgl. Frey & Hartig, 2022, S. 928; Gold et al., 2021, S. 109).

> „Dieser sich seit Mitte der 1990er Jahre vollziehende Wechsel des Fokus von den Voraussetzungen von Bildung auf ihre Ergebnisse, von Input- zur Outputsteuerung, geht mit einer zunehmenden Bedeutung der empirischen Kontrolle einher und kann als Paradigmenwechsel in der Steuerung des Bildungssystems bezeichnet werden“ (Schwippert & Coy, 2008, S. 392).

Kompetenzen lassen sich „als erlernbare kontext- und domänenspezifische kognitive Leistungsdispositionen“ (Frey & Hartig., 2022, S. 928 f.) definieren und demnach von

„Intelligenz als allgemeiner kognitiver Leistungsfähigkeit“ (Frey & Hartig., 2022, S. 928 f.) abgrenzen.

> „Kompetenzen, ihre Aneignung sowie ihre Entwicklung stellen in den großen Schulleistungsstudien (PISA, TIMSS, IGLU, NEPS etc.) einen, wenn nicht den zentralen Fokus der Untersuchung und Analyse von Lernbedingungen und Lernerfolgen dar“ (Bayer, 2015, S. 3).

Kompetentes Handeln bedeutet nicht primär den Besitz trägen Wissens, sondern die Fähigkeit immer wieder neue Anforderungssituationen bewältigen zu können (vgl. Bayer, 2015, S. 3).

> „Dieses Verständnis von Kompetenz als erlernbare, bereichsspezifische [und lebensphasenspezifische] Spezialkompetenz und als kognitives Dispositionskonstrukt [...] hat sich innerhalb der quantitativ-empirischen Kompetenzforschung durchgesetzt“ (Bayer, 2015, S. 3 f.).

Im Jahr 2003 wurden in Deutschland Bildungsstandards von der Kultusministerkonferenz eingeführt, welche Fähigkeiten und Kenntnisse beschreiben, die Schüler:innen am Ende der Primar- und Sekundarstufe erworben haben sollen. Zur Feststellung, inwieweit die in den Bildungsstandards formulierten Ziele von Schüler:innen erreicht wurden, dienen standardisierte Testverfahren, welche in regelmäßigen Intervallen durchgeführt und empirisch ausgewertet werden und individuelle Kompetenzausprägungen der Schüler:innen abbilden. Diesbezüglich wurde 2004 das Institut zur Qualitätsentwicklung im Bildungswesen an der Humboldt-Universität zu Berlin eingerichtet, um als zentrale Einrichtung (Leistungstest-) Verfahren zur Überprüfung der KMK-Bildungsstandards auf Bundesebene zu entwickeln (vgl. Fürstenau & Gomolla, 2012, S. 116).

2.3 Lernverlaufsdiagnostik als aktuelle Entwicklungstendenz

Derartige testdiagnostische standardisierte Verfahren standen häufig in der Kritik, da sie den Fokus weg von der Individualdiagnostik lenken und somit keinen Förderungs- und Inklusionszweck verfolgen, sondern vielmehr das Ziel einer leistungsbezogenen Klassifikation haben. Gold, Gawrilow und Hasselhorn beschreiben jedoch einen Trend in der schulpsychologischen Diagnostik, welcher sich von dieser Problematik distanziert:

> „Moderne schulpsychologische Diagnostik zielt stets auf die Unterstützung und Optimierung individueller Förderung und sie nimmt dabei die individuellen Lernverläufe in Relation zu den jeweiligen Lehr-Lern-Gelegenheiten in den Blick. Sie beschränkt sich nicht (mehr) nur auf die Erfassung allgemeiner Fähigkeiten und schulischer Kompetenzen, sondern nutzt differenzierte Möglichkeiten zur Feststellung der Funktionstüchtigkeit kognitiver Prozesse und Funktionen“ (Gold et al., 2021, S. 113).

Der aktuelle Trend, welcher hier beschrieben wird, ist der Wandel von der Lernstands- zur Lernverlaufsdiagnostik (LVD).

> „Lernverlaufsdiagnostik soll mithilfe wiederholter Testungen den individuellen Fortschritt, das Stagnieren oder das Abfallen von Leistungen dokumentieren, damit auf diese Prozesse kurzfristig pädagogisch reagiert werden kann“ (Gold et al., 2021, S. 115).

Es handelt sich um ein formatives Verfahren, da die Diagnostik der Anpassung und Planung des Unterrichts sowie von Fördermaßnahmen dient (Souvignier et al., 2021, S. 129). Inwieweit formative Assessments für die Anpassung der Unterrichtspraxis tatsächlich umgesetzt werden, bleibt jedoch unklar.

2.4 Formatives Assessment

Der Formative Diagnostik zielt „unmittelbar auf die Lehr-Lernprozesse“, so dass

Lernende eine „Rückmeldung darüber erhalten, wo sie gerade stehen um ihr weiteres Lernverhalten zu optimieren“ (Gold et al., 2021, S. 115). Gleichsam erhalten Lehrende eine Rückmeldung über den Erfolg ihres bisherigen didaktischen Vorgehens und können Lernangebote modifizieren (vgl. Gold et al., 2021, S. 115). An dieser Stelle erweist sich formative Diagnostik als große Chance. Anstatt eine summative Lernstandsdiagnose zu erstellen, ermöglichen serielle Testungen die Betrachtungen von Lernverläufen.

Die Ergebnisse standardisierter Tests können summativ und formativ genutzt werden (vgl. Schmidt, 2020, S. 9, S. 21). Häufig stehen groß angelegte standardisierte Leistungstests in Form von Leistungsvergleichsuntersuchungen in Form von ‚large-scale assessments' im Fokus. Summativ interessiert vorwiegend der Vergleich von Bildungssystemen unterschiedlicher Länder beziehungsweise Schulvergleiche mit dem Ziel des System- und Bildungsmonitorings (vgl. Schwippert & Coy, 2008, S. 390). Standardisierte Tests lassen sich jedoch auch für eine formative Diagnostik nutzen, in dem Lernverläufe betrachtet und für den Einsatz im Unterricht genutzt werden. Gegenstand dieser systematischen Literaturrecherche soll es daher sein, herauszufinden, welche Forschungsergebnisse es zur formativen Anwendung standardisierter Tests im Unterricht zur Unterstützung des Lernens gibt.

Insofern gilt es an dieser Stelle das begriffliche Konzept zum formativen Assessment zu klären, auf welchem die systematische Recherche aufbaut.

Das formative Assessment (FA) repräsentiert erst seit wenigen Jahren einen Themenbereich der Unterrichts- und Schulforschung im deutschsprachigen Raum. Der Begriff wurde ursprünglich im anglo-amerikanischen Raum im Rahmen des seit den 1960er Jahren existierenden Konzepts der summativen und formativen Evaluation genutzt. In Abgrenzung zum Begriff der Diagnostik, bei dem es sich in der Pädagogik und Psychologie um eine ‚Feststellung' handelt, richtet die Verwendung des Begriffs Assessment die Betrachtung eher auf die Wahrnehmung und Interpretation der Lehrperson und kann daher mit den deutschen Begrifflichkeiten ‚Deutung' oder ‚Einschätzung' beschrieben werden. Formativ meint im Deutschen ‚gestaltend' und bezieht sich demzufolge auf das Anliegen, auf Grundlage gewonnener Informationen über die Leistungen der Schüler:innen, durch das Adaptieren des Unterrichts den Lernprozess zu optimieren. (vgl. Black & Wiliam, 1998, S. 53; Schmidt, 2020, S. 7 f.)

Problematisch bei der Findung einer klaren Definition von formativem Assessment ist die begriffliche Unschärfe und die wissenschaftliche Uneinigkeit dieses Konzepts. Daher beziehe ich mich bei der Definitionsfindung für meine Recherche im Speziellen auf Schriften und das Review von Black und Wiliam (1998, 2018; Schmidt, 2020). Diese formulierten FA 1998 ursprünglich als Gegenentwurf zu bereits existierenden Praktiken zur Steigerung von Schüler:innenleistungen wie beispielsweise nationale Tests oder landesweite Vergleichstests, die mangels der Einbeziehung der Unterrichtsgestaltung in ihre Untersuchung nur eingeschränkt effektiv waren (vgl. Black & Wiliam, 1998, S. 54; Schmidt, 2020, S. 17). Black und Wiliam definieren als FA alle Aktivitäten, die Lehrpersonen in Interaktion mit den Schüler:innen durchführen, um Informationen über den Lernprozess der Schüler:innen zu erhalten und als Feedback zur Modi-fikation ihrer Lehr-Lern-Aktivitäten nutzen (vgl. Black & Wiliam, 1998, S. 7 f., 82). Es gibt verschiedene verwandte Begrifflichkeiten, welche fälschlicherweise synonym für das formativen Assessment eingesetzt werden, obwohl sie teilweise einen anderen Schwerpunkt haben, beispielsweise das Assessment for Learning, die formative Leistungsbeurteilung und die formative Leistungsdiagnostik. Während diese Begriffe vordergründig die Intention des Assessments meinen, setzt das formativen Assessment, wie es Black und Wiliam beschreiben, voraus, dass die Ergebnisse des Assessments auch tatsächlich genutzt werden, um den Unterricht kontinuierlich an die Lernbedürfnisse der Schüler:innen anzupassen, weshalb FA als fortwährender zyklischer Prozess zu verstehen ist (vgl. Schmidt, 2020,

S. 11, S. 21). „Der formative Assessment-Prozess beinhaltet […] die Schritte Ziele, Diagnose, Interpretation bzw. Beurteilung und Adaption der Unterrichtsaktivitäten“ (Schmidt, 2020, S. 14). Die OECD betrachtet die lernprozessbegleitende Diagnostik als Kernelement des FA, weshalb es auch als Variante von Lernprozessdiagnostik, die über einen kürzeren Zeitraum erfolgt, betrachtet werden kann (vgl. Schmidt, 2020, S. 14).

> „Zur Konkretisierung des Begriffs differenziert Wiliam (2009, 2010) zwischen long-cycle, medium-cycle und short-cycle formative Assessment. Zu long-cycle Assessment zählt laut Wiliam (2009) beispielsweise die Nutzung der Ergebnisse von jährlich stattfindenden nationalen Vergleichstests zur Ableitung von Fehlerschwerpunkten. Medium-cycle Assessment bezieht sich auf einen Zeitraum zwischen einer und vier Wochen. Allerdings beschreibt Wiliam beide Formen als wenig effektiv. Der größte Einfluss auf die Schülerleistung wird dem short-cycle Assessment (minute-to-minute, day-by-day) beigemessen“ (Schmidt, 2020, S. 12).

In ihrem Review (1998) kamen Black und Wiliam außerdem zu der Erkenntnis, dass eine häufige Durchführung von Leistungstests in den betrachteten Studien eine Leistungsverbesserung förderte. Dieser positive Effekt konnte jedoch nur festgestellt werden, so lange nicht häufiger als ein- bis zweimal pro Woche getestet wurde. Eine zu häufige Testung konterkarierte diesen Effekt (vgl. Black & Wiliam, 1998, S. 35). Der pädagogische Erfolg den Wiliam und Black mit dem Modell des formativen Assessment intendieren, wurde durch die Metaanalyse von Kingston und Nash (2011) geschmälert, da die Forscher im Abgleich von 300 Studien nur moderate Effektstärken (d = 0,25) dieser Intervention feststellen konnten. Außerdem lieferten nur 13 der betrachteten Studien qualitativ hochwertige Forschungsdaten. Dennoch wurde festgehalten, dass Feedback einen zentralen Wirkmechanismus von FA darstellt und es scheinbar besonders im Sprachunterricht effektiv ist. Formatives Feedback intendiert die Diskrepanz zwischen der aktuellen Leistung der Schüler:innen und den angestrebten Lernzielen zu reduzieren. Formativen Assessment soll in dieser Hinsicht metakognitive Strategien fördern, wie persönliche Zielplanung und Selbstassessment durch die Schler:innen, wodurch selbstreguliertes Lernen unterstützt wird (vgl. Schmidt, 2020, 33 f.; Havnes et al., 2012, S. 21). Man unterscheidet zwischen formellem formativem Assessment, das einer vorherigen Planung (im Unterrichtskontext üblicherweise durch die Lehrkraft) bedarf mit dem Ziel präzisere Informationen über die Lern- und Kompetenzentwicklung der Schüler:innen zu sammeln und dem informellen formativen Assessment, welches spontan innerhalb von Lehrer-Schüler- und Schüler-Schüler-Interaktionen erfolgt. Testbasierte Instrumente gelten mit ihrem hohen Standardisierungsgrad als formelle diagnostische Verfahren. Jedoch stellt sich die Anwendung standardisierter Testverfahren in der unterrichtlichen Praxis aktuell noch als problematisch dar. Als Gründe dafür benennt Schmidt (2020, S. 37), dass es häufig nur zwei Parallelformen solcher Tests gibt, was eine serielle Testung erschwert und daher die Beobachtung eines Lernverlaufs behindert. Außerdem schätzen Lehrpersonen den Einsatz solcher Tests im Unterricht als zu zeitintensiv ein und befürchten, deshalb zu wenig Zeit für die Umsetzung der Lehrplaninhalte zu haben (vgl. Schmidt, 2020, S. 37). Weiterhin fehlen den Lehrkräften Kenntnisse zur Entwicklung, Umsetzung und Auswertung solcher Testverfahren, also jenen Tätigkeiten die normalerweise durch Forscher*innen und Expert*innen erfolgen. Ebenso fehlen den Lehrkräften schulpsychologische diagnostische Grundlagen. Beides ist grundsätzlich nicht Teil der Lehrerausbildung (vgl. Schmidt, 2020, S. 55, S. 238; Souvignier et al., 2021, S. 136). Aus diesem Grund werden zunehmend computerbasierte formative Systeme entwickelt (vgl. Souvignier et al., 2021, S. 134). In den vorherigen Abschnitten wurde bereits thematisiert, dass die begriffliche Trennung von formativem Assessment, formativer Diagnostik und Lernverlaufsdiagnostik aufgrund der wissenschaftlichen Uneinigkeit über konkrete Definitionen und Unterscheidungskriterien diffizil ist. Grundsätzlich liegt

das Interesse dieser Arbeit auf der Betrachtung von standardisierten Leistungstests im Zusammenhang mit formativem Assessment, welches die Nutzung durch Lehrkräfte im Unterricht impliziert. Um ein umfangreiches und aussagekräftiges Rechercheergebnis zu erzielen, ist es jedoch notwendig auch andere der benannten Aspekte in die Suche einzubeziehen, die sich allgemein mit der formativen Nutzung standardisierter Leistungstests beschäftigen. Somit besteht auch eine Notwendigkeit in der Betrachtung, wie die jeweiligen Rechercheergebnisse den Aspekt des „Formativen" definieren und einordnen.

Zusammenfassend ergeben sich die folgenden Fragestellungen:
F1. Welche Forschungsergebnisse zur formativen Anwendung standardisierter Tests gibt es, die von Lehrpersonen (im Unterricht) zur Aufbereitung von Lernangeboten genutzt werden können?

F2. Welche Bedingungen braucht es, damit standardisierte Leistungstests ‚optimal' zur formativen Unterstützung im Unterricht genutzt werden können?

Zur Erörterung der Fragestellungen soll der aktuelle Forschungsstand zum Thema formative Nutzung standardisierter Leistungstests im Unterricht systematisch ausgearbeitet und die zentralen Ergebnisse zusammengefasst werden. Anschließend erfolgt, in Bezug auf die Fragestellung, ein Vergleich beziehungsweise eine Gegenüberstellung der Rechercheergebnisse.

3 Methodik

Diese systematische Literaturrecherche kombiniert Vorgehensweisen und Vorgaben, welche in den Leitlinienpapieren von Alexander (2020), Willems (2020) und Heil (2020) dargelegt werden. „Ein *Systematic Literature Review* ist eine eigenständige wissenschaftliche Methode mit dem Ziel, sämtliche relevante Literatur zum Forschungsthema zu identifizieren und kritisch zu bewerten" (Heil, 2020, S. 5). Die Recherche erfolgt dabei strukturiert und systematisch basierend auf festgelegten Suchkriterien, um aktuelles Wissen, in der Regel in Form von empirischen Forschungsbefunden, zu mindestens einer wissenschaftlichen Fragestellung zusammenzuführen (vgl. Willems, 2020). Durch eine detaillierte Recherchedokumentation, ist das Vorgehen während des systematischen Reviews transparent und reproduzierbar (vgl. Heil, 2020, S. 5).

Es wurde eine systematische Datenbankrecherche in den Datenbanken *PSYNDEX-plus*, *FIS Bildung*, *ERIC* und *Academic Search Elite* vorgenommen. Die Auswahl der Datenbanken erfolgte mithilfe der Kurzprofile, zusammengestellt von Hofmann (2013, S. 68 ff.) und dem Datenbank-Infosystem (DBIS) der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB).[2] Die durch Anwendung der im Folgenden definierten Ein- und Ausschlusskriterien ermittelten Publikationen wurden einem mehrstufigen Screening auf Passung unterzogen, um die zur Beantwortung der Fragestellung relevanten Forschungsberichte beziehungsweise Studien zu identifizieren.

3.1 Ein- und Ausschlusskriterien für Literatur

Um einen möglichst aktuellen Überblick zur formativen Nutzung standardisierter Tests, beziehungsweise im Zusammenhang mit formativem Assessment, zu bekommen, sollen lediglich Publikationen berücksichtigt werden, welche in den letzten fünf Jahren erschienen sind, also im Zeitraum zwischen 2017 und 2022. Aufgrund der eingeschränkten Datenlage zu diesem Thema, wurden sowohl Publikationen eingeschlossen, die sich mit

[2] Eine reproduzierbare Beschreibung des Suchverfahrens kann bei der Autorin eingeholt werden.

einer wissenschaftlichen Nutzung befassen, als auch solche, die Erfahrungen von Lehrpersonen und Schüler:innen einbeziehen. Außerdem konnten nur internationale Studien und Berichte einbezogen werden, die in deutscher oder englischer Sprache veröffentlicht wurden. Reviews, Kommentare, Essays, theoretische Schriften und Praxisbeschreibungen wurden nicht in Betrachtung genommen, da sie für die Beantwortung der Fragestellung ungeeignete Publikationsformen darstellen. Die Auswahl beschränkt sich weiterhin auf die Auseinandersetzung mit der Testnutzung im primären und sekundären, also weiterführenden, Regelschulkontext (z.B. Grundschule, Oberschule und Gymnasium) und schließt folglich die Vorschulausbildung sowie post-sekundäre Ausbildung aus. Um eine Übereinstimmung mit der Fragestellung zu gewährleisten, erfüllen nur Publikationen die Einschlusskriterien, welche konsistent mit den Definitionen eingangs beschriebener Begrifflichkeiten und Äquivalenten zu standardisierten Leistungstests, formativer Diagnostik beziehungsweise formativem Assessment sind. Zur Sicherstellung der Qualität der einzuschließenden Studien, wurde beim Screening darauf geachtet, dass die Publikationen möglichst begutachtet (peer review) sind und die Methodik der Studien klar ersichtlich war.

3.2 Literatursuche und Publikationswahl

Die Literatursuche wurde mithilfe von vier Datenbanken durchgeführt: FIS Bildung, ERIC, PSYN-DEXplus und Academic Search Elite. Die folgende Tabelle stellt die Suchbegriffe dar, welche in unterschiedlicher Weise und Kombination mit Operatoren verbunden respektive in verschiedenen Such-feldern genutzt wurden. Die Suchstrings variierten je nach Datenbank (vgl. Hofmann, 2013, S. 68 ff.). Die Begriffszuordnungen wurden mithilfe des Thesaurus von PSYNDEXplus vollzogen.

Tabelle 1: Suchbegriffe Literaturrecherche

Englisch	Deutsch
Formative Assessment	Formative Diagnostik verwandt: Feedback
Curriculum-based Assessment/ Measurement	verwandt: Lernverlaufsdiagnostik verwandt: formatives Assessment
Educational Measurement	allgemeiner: Pädagogisches Testen spezifischer: Lernzielorientierte Diagnostik
Standardized (Performance/ Achievement/ Proficiency) Tests	Standardisierte (Leistungs-) Tests
Educational Standards	verwandt: Bildungsstandards
Test Standardization	verwandt: Teststandardisierung
Teaching = classroom instruction/ lesson/ class	Unterricht verwandt: unterrichten = lehren
School	Schule

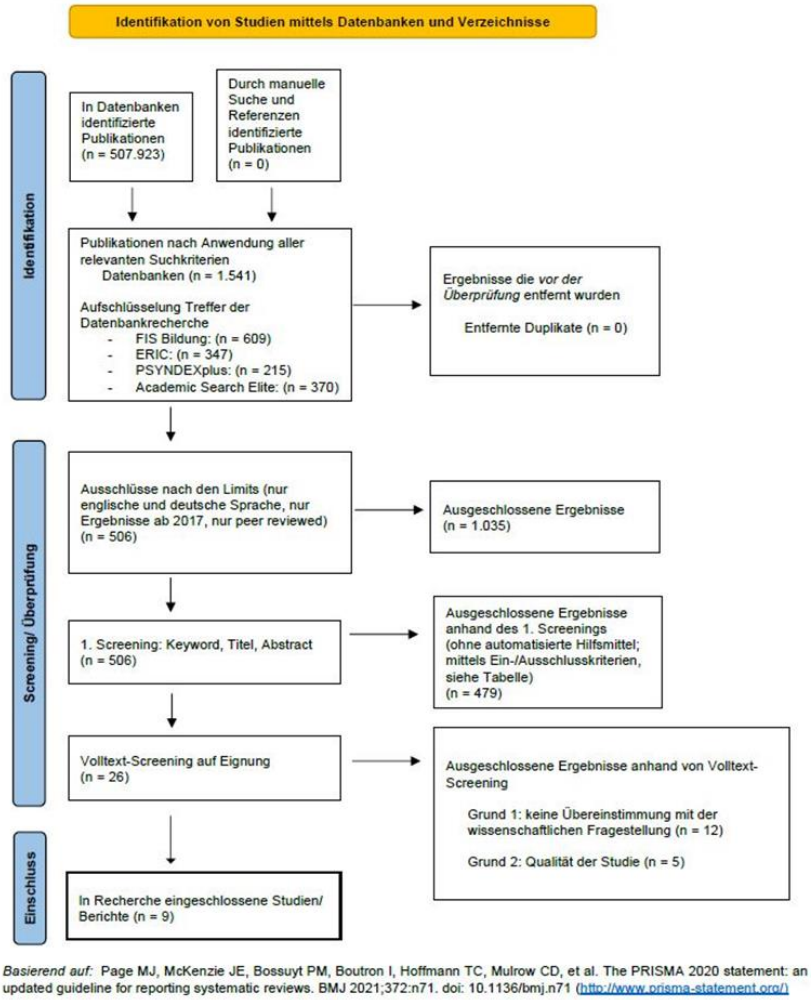

Abb 1: PRISMA-Flussdiagramm der systematischen Literaturrecherche

In den Datenbanken PSYNDEXplus, ERIC und Academic Search Elite wurden lediglich englische Suchbegriffe verwendet. In der Datenbank FIS Bildung konnte dagegen in englischer und deutscher Sprache gesucht werden. Aufgrund der großen Trefferzahlen (507.923 Suchergebnisse insgesamt), wurden nur Ergebnisse überprüft, die Treffer bei der Kombination der Begriffe zum Standardisierten Testen (Kriterium 1) in Verbindung mit ‚formativ‘/‚formative‘ (Kriterium 2) und gegebenenfalls ‚Schule‘/‚school‘ oder ‚Unterricht‘/‚teaching‘/‚classroom instruction‘ (Kriterium 3) ergaben. Teilweise erschienen Suchergebnisse lediglich durch Kombination von Begriffen, die sich auf die ersten beiden Kriterien beziehen. Durch das Entwickeln komplexerer Suchstrings mit Begriffskombinationen reduzierte sich die Anzahl der zu überprüfenden Publikationen auf 1.541.

Anschließend wurden die Einschränkungen verfeinert, indem nur Ergebnisse in die Suche eingeschlossen wurden, die begutachtet wurden (peer reviewed) und ab 2017 erschienen sind. Die Anzahl der Suchergebnisse reduzierte sich im Zuge dessen auf 506. Sofern die Jahreszahl nicht als Filter eingestellt werden konnte (beispielsweise bei FIS Bildung), wurden Publikationen, die vor 2017 erschienen sind, im ersten Screening (des Titels, der Keywords, ggf. des Abstracts), manuell ausgeschlossen. Suchergebnisse, die nach Überprüfung der Keywords und des Titels als passend er-

schienen, wurden weiter auf die Übereinstimmung mit der wissenschaftlichen Fragestellung beziehungsweise den Einschlusskriterien (Kap. 3.1) durch das Screening der jeweiligen Abstracts überprüft. Sofern die gefundenen Publikationen online auffindbar beziehungsweise der Zugriff darauf möglich war, wurden sie schließlich in das Volltext-Screening aufgenommen. Auf diese Weise wurden 26 Studien und Berichte ausgewählt, deren Eignung durch Überprüfung der Volltexte erfolgte. Nach Analyse der Volltexte wurden 17 weitere Studien ausgeschlossen, darunter zwölf Publikationen, deren Untersuchungen nicht unmittelbar zur Fragestellung dieser Arbeit passten und Fünf weitere, die nicht den in den Einschlusskriterien benannten Qualitätskriterien entsprachen. Die Anzahl der eingeschlossenen Studien für die systematische Literaturrecherche reduzierte sich dadurch auf neun. Im Folgenden wird der soeben beschriebe-ne Ablauf der Identifikation von Studien für die systematische Recherche durch ein sogenanntes PRISMA-Flussdiagramm illustriert. Es handelt sich dabei um ein von Moher et al. (2009) entwickeltes System zur Dokumentation systematischer Reviews. Das Diagramm basiert auf den nachstehenden Richtlinien (vgl. Moher et al., 2009; Page et al., 2021).

3.3 Eingeschlossene Studien bzw. Forschungsberichte

Die folgenden neun Publikationen wurden in diesem Review untersucht:

Tab 2: Eingeschlossene Publikationen

1	Alexander et al., 2020
2	Edmentum Research, 2018a
3	Edmentum Research, 2018b
4	Faber & Visscher, 2018
5	Karagiorgi & Petridou, 2019
6	Manyak & Manyak, 2021
7	Ponce et al., 2018
8	Sutter et al., 2020
9	Zheng et al., 2019

Bei den eingeschlossenen Studien finden verschiedene Studiendesigns Anwendung. Die Studien eins bis drei sind Beobachtungsstudien und bei den Studien vier sowie sechs bis neun handelt es sich um Experimentalstudien[3]. Fast alle Studien nutzen sowohl quantitative als auch qualitative Daten für ihre Analysen. Publikation fünf stellt eine Ausnahme dar, da es sich dabei um einen Fallbericht handelt, der ein staatlich organisiertes wissenschaftliches Programm zur Früherkennung von Schulversagen mittels formativen Einsatzes von standardisierten Leistungstests beschreibt, diskutiert und wissenschaftlich

[3] Teilweise wird in den Studien das jeweilige Design nicht deutlich definiert. Die Einschätzung bzw. Zuordnung zu einem Studiendesign erfolgte dann nach bestem Wissen und Gewissen basierend auf der Methodik und Datenlage der jeweiligen Publikation.

belegt. Zwar können hierbei keine Angaben zur Studienqualität gemacht werden, weil dazu keine ausreichenden Daten vorliegen. Dennoch erachte ich den Artikel hinsichtlich der Erörterung des Forschungsstandes zur formativen Nutzung standardisierter Tests für relevant. In den Publikationen sind sprachliche (vgl. Edmentum Research, 2018a, 2018b; Ponce et al., 2018; Faber & Visscher, 2018; Manyak & Manyak, 2021; Sutter et al., 2020; Karagiorgi & Petridou, 2019), mathematische (vgl. Edmentum Research, 2018a, 2018b; Karagiorgi & Petridou, 2019; Zheng et al., 2019) und naturwissenschaftliche (vgl. Edmentum Research, 2018a, 2018b; Alexander et al., 2020) Fächer und Kompetenzen Untersuchungsgegenstand formativer Tests und Interventionen. Das bestätigt den eingangs geschilderten Fokus standardisierter Leistungsuntersuchungen auf schulische Kernfächer, welche gesellschaftliche Grundfähigkeiten und Kulturtechniken vermitteln und gleichzeitig konvergente, leicht messbare Leistungen abbilden (vgl. Jürgens, 2022, S. 553; Schwippert & Coy, 2008, S. 398).

Die Studien und der Bericht stammen aus den USA, Zypern, Chile und den Niederlanden. Folglich wurden die Studien vor dem Hintergrund differenter Bildungssysteme durchgeführt. Es liegen daher beispielsweise unterschiedliche schulische Selektionsmechanismen zugrunde. Man kann deshalb nicht voraussetzen, dass die Schüler:innen nach der Grundschulzeit nach Leistung verschiedenen Schulformen zugewiesen werden, wie im deutschen Bildungssystem. Schulformen und Fächer können sich außerdem unterscheiden, weshalb man von der Schulform auch nicht auf ein Leistungsniveau schließen kann, wie es im deutschen Bildungssystem mit der Unterscheidung von Oberschule und Gymnasium üblich ist. Daher wird in dieser Arbeit, bei Betrachtung der Studienergebnisse, nur in Altersgruppen unabhängig von der Schulform unterschieden. Betrachtet man nun die Klassenstufen eins bis sechs als Grundstufe, die Klassenstufen sieben bis neun/zehn als Mittelstufe (vergleichbar mit Oberschule, Middle School oder Gymnasium) und die Klassenstufen elf bis zwölf als Oberstufe (vergleichbar mit Gymnasium oder High School), so ist hervorzuheben, dass sich beinahe die Hälfte der Publikationen mit dem Grundschul- respektive Elementarniveau beschäftigen. Aus den Studien geht als Grund für diesen Fokus unter anderem die frühzeitige Diagnostik von Lern- und Leistungsschwäche und einer damit ggf. einhergehenden Schullaufbahngefährdung, durch formative beziehungsweise standardisierte Testung hervor (vgl. Faber & Visscher, 2018; Karagiorgi & Petridou, 2019; Manyak & Manyak, 2021; Sutter et al., 2020). Nur eine Studie beschäftigt sich mit der Mittelstufe (vgl. Zheng et al., 2019). Auch auf der Oberstufe liegt ein besonderer Schwerpunkt für die Betrachtung der Nutzung von Leistungstests, weil solche Testungen in dieser Altersstufe im Speziellen im Zusammenhang mit der Vergabe von Abschlusszertifikaten relevant sind (vgl. Alexander et al., 2020; Ponce et al., 2018). Das Software-Unternehmen Edmentum verfolgt das Interesse, Programme zur Durchführung von formativem Assessment und Tests für alle Altersgruppen bereitzustellen. Daher liegt in dieser Studie das Augenmerk auf allen Schulstufen (vgl. Edmentum Research, 2018a, 2018b). Generell ist aus den Publikationen ein Trend in Richtung Online-Testung oder auch der Entwicklung und Nutzung von digitaler formativ einsetzbarer Assessment-Software zu verzeichnen. Hauptsächlich dienen klassische, zumeist staatliche, standardisierte Leistungstests in diesem Zusammenhang der Erprobung der Effektivität solcher Programme mittels statistischer Regressionsanalysen der Ergebnisse aus den Programmen und aus den standardisierten Tests.

3.4 Datenextraktion und Qualität der Publikationen

Als Kriterien zur Beurteilung der Qualität quantitativer empirischer Studien benennt Döring (2015) die Güte und Vollständigkeit der Berichterstattung, also ob wissenschaftliche Standards, wie das Formulieren eines Forschungsproblems und die Aufarbeitung des Theorie- und Forschungsstandes, eingehalten wurden. Weitere diesbezügliche

Kriterien sind die Größe, Repräsentativität und Zufälligkeit der Stichprobe, die Operationalisierung der Studie durch geeignete Messinstrumente zur Datenerhebung, die transparente Datenanalyse und -interpretation durch geeignete statistische Verfahren sowie die Konstruktvalidität der Studien (vgl. Döring, 2015). Da es sich jedoch um unterschiedliche, nicht unbedingt vergleichbare Studiendesigns handelt, fehlt die Grundlage für eine umfassendere Bewertung der Studienqualität und -güte. Die Qualität der Publikationen wurde folglich dahingehend überprüft beziehungsweise bestätigt, dass sie bereits bei der Recherche nach dem Einschlusskriterium ausgewählt worden sind und ob sie peer-reviewed sind. Außerdem erfolgte die Suche ausschließlich über anerkannte wissenschaftliche Fachdatenbanken (sh. oben), wodurch auf eine wissenschaftliche Relevanz und Güte der ausgewählten Arbeiten geschlossen werden kann. Bis auf die Studien von Edmentum (2018a, 2018b) sind alle Publikationen in renommierten Fachzeitschriften erschienen. Grundsätzlich zeichnen sich alle Publikationen durch eine detaillierte Berichterstattung, Datenerhebung, -analyse und -interpretation sowie durch transparente Methodik und Theoriebezug aus, wie nachfolgend und in Tabelle 3 genauer beschrieben wird. Die Studien von Edmentum (vgl. Edmentum Research, 2018a, 2018b) hat den Mangel, dass keine Urheber*innen und Autor*innen der Studie benannt werden, es wird nur allgemein auf die gleichnamige Software-Firma verwiesen. Dennoch liefern beide Studien genaue Angaben zum Studieninhalt, der Methodik, Theorie und Daten zu Forschungsergebnissen, sodass diese aufgrund der inhaltlichen Relevanz eingeschlossen werden konnten. Bei der Publikation von Karagiorgi und Petridou (2019) handelt es sich um einen Fallbericht ohne detaillierte Forschungsdaten, der allerdings eine fundierte wissenschaftliche Berichterstattung zu erfolgten Studien enthält, welche ebenfalls die inhaltlichen Kriterien der Fragestellung erfüllt.

4 Ergebnisse

Im Folgenden l werden die einbezogenen beziehungsweise Artikel hinsichtlich der Forschungsfrage oder des -themas vorgestellt und der jeweilige Untersuchungsgegenstand erläutert. Anschließend werden die Ergebnisse der Studien vor dem Hintergrund der Nutzung standardisierter Leistungstests im formativen Kontext im Unterricht hinsichtlich ihrer Gemeinsamkeiten, Widersprüche, Betrachtungsweisen und Besonderheiten beschrieben

4.1 Beschreibung der Einzelpublikationen

Tabelle 3: Charakteristika und Inhalte der eingeschlossenen Publikationen

Studie Publikationsart	/Untersuchungs-gegenstand	Design /Methoden Kontext		Stichprobe	Eingesetzte Tests	Ergebnisse
1 Alexander et al., 2020 Forschungs-bericht	Nutzen von formativem Assessment in Vorbereitung auf summative stand. Tests	Beobachtungs-studie USA (Mississippi) Fach: Career and Technical Education (CTE)*	Datensammlung- und -vergleich: - MS-CPAS-Testergebnisse, Schuljahre 2018/19, vor und nach Einsatz des formativen Assessments - CTE-Tests	n=28.015 SuS (n= 54 CTE-Kurse) 10./11. Klasse freiwilliger Einsatz des CTE-Übungstests durch Lehrkräfte	MS-CPAS-Test:(Mississippi Career and Planning Assessment System, online): - summativer stand. Test - 100 MC-Fragen CTE-Test: - formativer online Test basierend auf Fragen aus früheren MS-CPAS-Tests, ca. 50 MC-Fragen	Sign. Unterschiede zwischen summativen Testergebnissen von Schüler:innen, die zwei formative Onlinetests und Schüler:innen, die keine Übungstests absolviert haben: Effektstärke: d=0,53 (mittel)
2 3 Edmentum Research, 2018a und 2018b Forschungs-bericht	Untersuchung des Zusammenhang von Nutzungsmenge und Leistungs-daten der Schüler:innen innerhalb von Study Island verglichen mit deren Abschneiden im PSSA	Beobach-tungsstudie USA (Pennsylvania) Fächer: - Englisch - Mathe - Science (Kl. 3-8,11)	Datenabgleich mittels Propensity Score-Matching nach Fähigkeitsniveaus → Vergleich der SuS-Ergebnisse des PSSA 2016 mit PSSA 2017	Schüler:innen aus 24 Schulen (14 GS, OS, 3 GY, 3 AS) PSSA-Testergebnisse aus den Jahren 2016 und 2017 Study-Island-Nutzungs- und Performance-Daten	Study Island: formatives computerisiertes Übungs-und Assessment-Tool mit direktem Feedback → *Übungstests* → *Benchmark-Tests* - automatisierte online-Auswertung und Rückmeldung direkt an L PSSA: (Pennsylvania System of School Assessment)	Nutzung variiert nach Fach und Klassenstufe - durchschnittlich geringe Nutzung durch Schüler:innen (wenig Zeit investiert, wenig Aufgaben bearbeitet) Sign. positiver Einfluss auf Leistung im PSSA durch Nutzung von Study Island - starker Zusammenhang zwischen Benchmark-Ergebnissen und PSSA-Ergebnissen → Koeffizient 0,57-0,85 → hohe Vorhersagevalidität Benchmark-Tests
4 Faber &	Untersuchung der	Experimentals	Experimental-	30 Experimental-	Digitales Formatives Assessment-	Kein Hinweis darauf, dass Assessment-

	Visscher, 2018 Forschungs-bericht	Effekte eines digitalen formativen Assessment-Tools auf die Rechtschreibung von Drittklässlern	tudie Land: Niederlande Fach: Niederländisch	schulen nutzten das Assessment-Tool, Kontrollschulen führten regulären UR durch	schulen (n=619), 39 Kontrollschulen (n=986) Daten: - Ergebnisse standardisierter Vor- und Nach-Tests (Cito-Test) 2014/15 - Schüler:innen-Umfragen	Tool (DFAT) „Snappet“: - an Bildungsstandards orientierte oder an Lernziel orientierte Übungen, adaptive Aufgaben, die sich nach individuellem Leistungstand richten - direktes richtig/falsch-Feedback an Schüler:innen, L können Fortschritt der Schüler:innen auf Dashboard mitverfolgen Cito-Test: standardisierter Orthografie- und Mathematiktest	Tool Erwerb von Fähigkeiten der Rechtschreibung beeinflusst: - Orthografie-Leistung der Experimentalgruppe nicht besser als von der Kontrollgruppe - geringfügiger Einfluss auf Leistung, je mehr Aufgaben erledigt wurden (r=0.09) bzw. je mehr adaptive Aufgaben erledigt wurden (r=0,11) - höherer Einfluss von Motivation auf Nachtest (r=0,19)
5	Karagiorgi & Petridou, 2019 Wissen-schaftlicher Artikel	Identifikation „gefährdeter Schüler:innen (Kl. 3 und 5) und Unterstützung mittels adaptiven UR	Beobachtungs bericht Land: Zypern Fächer: Griechisch, Mathematik	Leistungstests in Griechisch und Mathe (Klasse 3 und 6) Schüler:innen- und L-Fragebögen (u.a. Lerneinstellung, familiäres Umfeld)	KA	Nationales *Programme for Functional Literacy (PfL)*: - zwei standardisierte Tests, formativ durchgeführt in Klassen 3 und 6, um Schüler:innen mit Risiko auf Schulversagen zu identifizieren - MC- und CR-Items	→ 2011-2015, rund 7-13% der Schüler:innen sind „gefährdet“ in Griechisch/ Mathe → gefährdete Schüler:innen sind häufig männlich, mit Migrationshintergrund, wurden zeitiger eingeschult, haben eine negative Einstellung zu den untersuchten Fächern bzw. geringes Selbstvertrauen
6	Manyak & Manyak, 2021 Artikel/ Studien-bericht	Testung eines Vokabellern-programms in der dritten Klasse zur Erweiterung des Vokabulars und zum Erwerb von diesbezüglichen	Experi-mentaldesign Land: USA (Colorado) Fach: Sprach-unterricht	Interventionen von Vokabellernen über 3 Jahre hinweg) Leistungsdaten sowie drei Assessments der Schüler:innen (GMVT)	High Valley Elementary School (Pseudonym), K-5 n=250 Schüler:innen Zusammenarbeit mit einer Lehrerin	Vokabel-Subtest des Gates-MacGinities Reading Test (GMVT) - 45 MC-Items, Durchführung zu Beginn und am Ende des Schuljahres 3 Assessments (SVKA, CAA, MAA) zu Instruktion, jeweils zu Beginn und Ende des Schuljahres	Schüler:innen hatten nach der Intervention ein größeres Grundvokabular im Vergleich zur Normierungsstichprobe Ermittlung der Effektstärke mittels paarweisen t-Tests: - bei allen Tests hohe Effektstärke

		Lernstrategien	(fächerübergreifend)	Sammlung qual. Daten durch UR-Beobachtung		- hohe interne Konsistenz	→ GMVT: d=0,93-1,45 → SVKA: d=1,23-1,61 → CAA: d=0,87-1,36 → MAA: d=1,52-2,38
7	Ponce et al., 2018 Artikel/ Studien-bericht	Untersuchung des Effekts einer Software, welche FA nutzt, auf das Vokabellernen und Leseverständnis	Experi-mentalstudie Land: Chile Fach: Englisch	Softwarenutzung für 10 Wochen in der Experimental-gruppe (EG) - Schulung von L - Vergleich von EG und KG stand. Leseverstehen Test	N=99 Schüler:innen der 11. Klasse von zwei GY → Schule 1: n=30 Schüler:innen in KG n=19 Schüler:innen in EG → Schule2: n=29 Schüler:innen in KG n=21 Schüler:innen in EG	Leseverständnis-Test des National Test System for English (SIMCE) als Vor- und Nachtest → im Vortest niedrige Leistung in beiden Schulen - Papierform, 45 Minuten	Experimentalgruppe, die Intervention genutzt hat, schnitt im standardisierten Leistungstest nach dem Treatment besser ab als Kontrollgruppe mit einer sehr großen Effektstärke (d=1,60 in Schule 1, d=1,17 in Schule 2)
8	Sutter et al., 2020 Forschungs-bericht	Effekte formativen Lese-Assessment Identifikation von Leseschwierigkeiten im Grund-schulalter, Interventionen	Experi-mentalstudie Land: USA Fach: fächerüber-greifend	Überprüfung der Vorhersagekraft des ISIP-ER-Tests mittels STAR Reading-Assessment	Schüler:innen (n=428) der 2. Klasse aus Grundschulen	CAT = computeradaptiver Test, mit automatischer Auswertung Instation's Indicators of Progress for Early Reading (ISIP-ER), zur Erfassung von Lesekompetenzen, unmittelbare Information der L über Leistungsentwicklung STAR Reading-Assessment	Der ISIP-ER-Test ist vorhersagekräftig → Effektgröße $R^2 = 0,67$ die beiden Tests korrelieren stark, r = 0,81 ISIP-ER ist als Vorhersageinstrument für die Leseleistung nutzbar

9	Zheng et al., 2019 Forschungs-bericht	Überprüfung der Vorhersagekraft eines adaptiven formativen Tutoring-Systems	Experi-mental-studie Fach: Mathe Land: USA (Florida)	Schüler:innen der Stichprobe nutzten MATHia 4 Jahre lang (Nutzungsdaten) Hinzunahme der Ergebnisse aus FCAT bzw. FSA-Test (Nutzung als Vor- und Nachtest)	n=23,374 SuS auf Klassenstufen 7-8 im Schuldistrik Miami-Dade County Public Schools	MATHia: intelligentes Tutoring-System für den Mathe-UR - adaptives System, Schüler:innen erledigen „Workspaces“ (realitätsnahe, mehrschrittige Aufgaben) nach dem Konzept des „Mastery Learnings“ (Bloom) FCAT (Florida Comprehensive Assessment Test) und FSA-Test (Florida Standards Assessment)	Die besten Vorhersagen lieferte das statistische Modell, welches die meisten Daten inkludierte, Modell ohne demografische Daten auch valide dennoch können die mithilfe von MATHia-Daten ermittelten Vorhersagen die stand. Leistungstests in ihrer Aussagekraft nicht ersetzen

Anmerkung. Die Angaben entsprechen den Details, die in dieser Arbeit dargestellt werden. Es werden nur Teilnehmende und Aspekte in dieser Tabelle berücksichtigt, welche für die Forschungsfragen dieses Reviews relevant sind. Abkürzungen in dieser Tabelle: KA = keine Angabe, stand. = standardisiert, UR= Unterricht, L = Lehrperson, GS = Grundschule/ Elementary School, OS = Oberschule/ Middle School, GY = Gymnasium/ High School, AS= Alternative Schule, MC = Multiple-Choice, CR = Constructed-Response, SC = Naturwissenschaften (Sciences), ELA = English Language Arts.

4.2 Synopse

Die neun Publikationen beleuchten unterschiedliche Nutzungsmöglichkeiten standardisierter Tests im formativen Kontext. Hierzu konnten die Publikationen in drei Kategorien unterteilt werden:

1. Kategorie: Nutzung von formativem Assessment zur Vorbereitung auf summative standardisierte Leistungstests, beispielsweise am Jahresende (Alexander et al., 2020; Edmentum Research, 2018a, 2018b)

Insbesondere in den USA ist diese Thematik von Relevanz (6 der 9 eingeschlossenen Publikationen stammen aus den USA), da die Durchführung von standardisierten Leistungstests am Schuljahres-ende von großer Bedeutung im amerikanischen Schulsystem ist (vgl. Gold et al., 2021, S. 111). Die Ergebnisse solcher Tests bestimmen sowohl die Wettbewerbsfähigkeit als auch Rechenschaftslegung einzelner Schulen und Schuldistrikte (vgl. Alexander et al., 2020, S. 336; Sturm, 2016, S. 43-65). Ebenso dienen diese Tests bspw. prominent der Scholastic Assessment Test (SAT) der Vergabe von Abschlusszertifikaten und tragen somit zur Entscheidungsfindung über die Qualifikation für und Allokation von akademischen Laufbahnen bei (vgl. Alexander et al., 2020, S. 336; Sturm, 2016, S. 43-65).

In den drei Studien (vgl. Alexander et al., 2020; Edmentum Research, 2018a, 2018b) erhielten die Lehr-kräfte keine konkreten Vorgaben, wie und in welchem Umfang sie die formativen Tests in ihrem Unterricht einsetzen sollten.

> „It is certainly possible that some teachers administered the two separate practice tests in the few days leading up to the summative exam. This scenario would have likely eliminated the main benefit of a formative assessment, which is high-quality feedback given to the students“ (Alexander et al., 2020, S. 344).

Zudem wurde nicht beobachtet, wie die Lehrpersonen die Auswertungsdaten nutzten und inwiefern sie mit der Adaption ihres Unterrichts und Lehrer-Feedback auf individuelle Leistungen und Schwierigkeiten der Schüler:innen reagierten. Alexander et al. betonen hierzu, mit Bezug auf den bisherigen Forschungsstand, dass formatives Assessment nur in Verbindung mit direktem Feedback durch die Lehrkräfte effektiv ist. Onlinetestformate erleichtern das Geben von Feedback ungemein, da die Ergebnisse schnell und genau sowie automatisiert ausgewertet werden können. Demnach bieten sie die Chance, das Lehrerfeedback qualitativ zu bereichern (vgl. Alexander et al., 2020, S. 344 f).

> „Kulik and Kulik (1988) found that immediate feedback of results has a positive effect on student achievement within classroom settings, especially on applied learning measures such as frequent quizzes. […] Pollock (2001) further concluded that feedback that also provided an explanation of the correct answer was the most effective. […] Although most of the research literature has focused on the effect of teacher-provided feedback or feedback from classroom-based assessments, research has shown that computers are also effective tools for providing feedback“ (Edmentum Research, 2018a, S. 3).

So integriert das Study Island-Tool aus der Edmentum-Studie bereits computerisiertes direktes Feedback an die Schüler:innen. In beiden Edmentum-Studien wird resümiert, dass die Analyseergebnisse deutlich von der Qualität und der Vorgehensweise bei der Nutzung von Study Island durch die Schulen beeinflusst wurden. Daher wäre es zukünftig notwendig qualitative Nutzungsunterschiede zu beobachten und insofern herauszufinden, welche Anwendungsmethoden am ertragreichsten für den Lernprozess der Schüler:innen sind (vgl. Edmentum Research 2018a, S. 24; Edmentum Research 2018b, S. 28). Auch Alexander et al. schlagen für die zukünftige Forschung vor, dass die Nutzung von formativem Assessment im Onlineformat weiter untersucht wird und dass

Lehrkräfte besser in der Nutzung solcher digitalen Testwerkzeuge geschult werden sollten (2020, S. 345 f.). In Hinblick auf Feedback wird empfohlen, die Feedbackmethoden der Lehrpersonen aufgrund ihrer Relevanz im Rahmen des formativen Assessments mehr in die Untersuchungen einzubeziehen (Alexander et al., 2020, S. 345 f.).

2. Kategorie: Erprobung der Effektivität von formativen Übungs- und Assessment-Werkzeugen mithilfe von standardisierten Leistungstests; Untersuchung der Vorhersagekraft von formativen Assessment-Angeboten auf die Schüler:innenleistung in standardisierten Leistungstests (vgl. Faber & Visscher, 2018; Manyak & Manyak, 2021; Ponce et al., 2018; Sutter et al., 2018; Zheng et al., 2019)

Diese Kategorie beschreibt eine primär wissenschaftliche Nutzung standardisierter Leistungstests zur Effektivität formativer Assessments. Die Wirksamkeit formativer Treatments wird dabei mittels standardisierter Vor- und Nachtests und Korrelationsanalysen der Testergebnisse überprüft. Immer häufiger kommen dabei computerbasierte Assessments bzw. Tests zum Einsatz, die sich durch einen hohen Standardisierungsgrad auszeichnen und besonders für den unterrichtlichen Einsatz geeignet sind, weil sie automatisiertes Feedback geben sowie be- und auswerten und Lehrkräfte somit entlasten können.

Faber & Visscher, 2018:

Die Häufigkeit der Nutzung von Digitalen Formativen Assessment-Tools (DEFAT) nimmt zu. Laut Faber und Visscher wird die App Snappet zum Beispiel bereits an insgesamt 2000 von 6500 Grundschulen in den Niederlanden eingesetzt. In beiden Studien von Faber und Visscher (2016, 2018) konnte festgestellt werden, dass eine häufige Nutzung dieses Tools zu höheren Leistungsniveaus führte. Kingston und Nash stellten diesbezüglich in ihrer Meta-Analyse fest, dass sich formatives Assessment sowohl in komplexen Fächern (z.B. Mathematik, Naturwissenschaften) als auch weniger komplexen Fächern (z.B. Sprachen) als effektiv darstellt, wobei die Effektgrößen in weniger komplexen Fächern höher sind. „[T]he use of computer-based formative systems, appeared to be more effective than other approaches, yielding mean effect size of .30 and .28“ (Kingston & Nash, 2011, S. 28). Mit Bezug auf äußeres Feedback stellten Hellrung und Hartig jedoch fest, dass es für den Erfolg im Fach Mathematik einen größeren Effekt hat als beim Lesen, weil Lehrkräfte in solchen klarer strukturierten Fächern einfacher Veränderungen im Lernfortschritt der Schüler:innen feststellen können (vgl. Hellrung & Hartig, 2013). Auch in dieser Studie wurde nicht beobachtet, wie die Lehrkräfte ihren Unterricht mithilfe des Snappet-Tools angepasst haben. Die Ergebnisse machen deutlich, dass Fachspezifika sowie daran angebundene Feedbackformen in der Adaption formativen Assessments zu berücksichtigen sind. Faber und Visscher betonen die Relevanz qualitativer Komponenten in kommenden Untersuchungen, insbesondere wie Lehrkräfte und Schüler:innen das Feedback nutzen und welche Informationen DFAT-Tools besonders hilfreich für sie sind, um Lernen und Unterrichten zu verbessern. Außerdem sollte zukünftig auch die Effektivität von adaptiven Aufgaben im Vergleich zu statischen Aufgaben genauer überprüft werden (vgl. Faber & Visscher, 2018, S. 6 f.).

Manyak & Manyak, 2021:

In der Studie von Manyak und Manyak (2021) wurden standardisierte Leistungstests formativ als Vor- und Nachtest genutzt, um die Effektivität neuer Lehrmethoden bezogen auf den Kompetenzerwerb der Schüler:innen zu überprüfen. Hervorzuheben ist an dieser Stelle, dass ausschließlich in dieser Studie auch Hospitationen im Unterricht stattfanden, um die Reaktionen der Schüler:innen auf die Interventionen, wie sie neue Techniken und neues Wissen anwenden, zu beobachten. Weiterhin wurde die Interaktion zwischen der Lehrkraft und den Schüler:innen genauso betrachtet, wie formale Aspekte

der Umsetzung, z.B. die zeitliche Komponente, welche beschreibt, wie lange einzelne Unterrichtsphasen dauerten. Bezogen auf den Unterricht zeigt die Studie, dass man mithilfe von standardisierten Leistungstests die Effektivität von Lehrmethoden und Unterrichtsreihen überprüfen kann. Jedoch fehlt es Lehrkräften an der statistischen Kompetenz, derartige valide Tests mit einer hohen kriteriellen Güte selbst zu entwickeln. Grundsätzlich sei das nur in Zusammenarbeit mit Expert:innen möglich.

Ponce et al., 2018:

In der Studie von Ponce et al. wurde ein standardisierter Leistungstest als Vor- und Nachtest durchgeführt, um die Effektivität einer Software für formatives Assessment in Echtzeit nachzuweisen. Damit wird bestätigt, dass formatives Assessment in Echtzeit mit direktem Feedback an Lehrkräfte und Schüler:innen einen sehr hohen Einfluss auf den Lernzuwachs der Schüler:innen hat, wobei die unterrichtlichen Aktivitäten unmittelbar an die Bedürfnisse der SuS angepasst werden. Problematisch sehen die Wissenschaftler:innen die aktive Nutzung der Software durch die Lehrkräfte, welche mit Unterrichtsadaptionen verbunden ist. Ponce et al. (2018) beschreiben diesbezüglich Unsicherheiten auf Seiten der Lehrkräfte, ihren traditionellen Unterrichtsstil anzupassen. Bei der Entwicklung und dem Einsatz neuer formativer Assessment-Methoden ist demzufolge eine enge Kooperation zwischen Entwickler:innen, Wissenschaftler:innen und Lehrkräften notwendig genauso wie Schulungen des Lehrpersonals, um einen hohen Ertrag der Methoden zu erzielen (vgl. Ponce et al., 2018, S. 57 f.).

Sutter et al., 2018:

Die Studie von Sutter et al. rückt den Fokus insbesondere auf den zunehmenden Einsatz von Computeradaptiven Tests (CAT) als Hilfsmittel für Leistungs-Screening und Benchmarking. Auch Sutter et al. betonen die Nutzung derartiger Testergebnisse für Unterrichtsadaptionen und dass Lehrkräfte in der Nutzung von solchen Testdaten für ihre Unterrichtsgestaltung dementsprechend geschult werden: „To maximize instructional decision making, professional teacher development that targets making meaning of learning analytics for instructional decisions should be implemented“ (Sutter et al., 2018, S. 50). Motivationale und situationale Faktoren wurden nicht in die Studie aufgenommen. Sutter et al. erachten den Vergleich von CATs und den klassischen schriftlichen, den sogenannten Paper-Pencil-Tests, in zukünftigen Untersuchungen für notwendig. Dabei sollte auch erkundet werden, ob es einen Unterschied in der Vorhersagekraft von schriftlichen und computerisierten Tests gibt.

Zheng et al., 2019:

In den USA nehmen summative Tests einen großen Anteil der Unterrichtszeit ein. In dem Schuldistrikt, in dem die Studie durchgeführt wurde, entfallen circa 20% der Mathematik-Unterrichtszeit auf die Vorbereitung, Durchführung und Auswertung standardisierter Tests (vgl. Zheng et al., 2019, S. 154).

> „This opportunity to rethink the relationship between formative and summative assessment comes at a time of increasing concern about current approaches to summative assessment and school accountability systems […]. Some of these concerns reflect the disconnect between assessment and instruction“ (Zheng et al., 2019, S. 154).

Zheng et al. befürworten die Entwicklung von im Unterricht inkludierten formativen Assessment-Möglichkeiten, die ebenso wie summative Tests den Lernprozess der Schüler:innen mitverfolgen und Rückschlüsse für die Unterrichtsgestaltung geben. „Our argument, then, is not that testing cannot be a valuable educational experience, but that summative testing is not designed to be educational“ (Zheng et al., 2019, S. 155). Weiterhin argumentieren sie, dass klassische (also summative) standardisierte Leistungstests weniger Einblick in vorhandenes Wissen der Schüler:innen geben als im Unterricht eingebautes (formatives) Assessment (vgl. Zheng et al., 2019, S. 170). Zheng

et al. plädieren für die Ersetzung formaler summativer Leistungstests durch formative Assessment-Strategien.

3. Kategorie: Formative Nutzung standardisierter Leistungstests zur Früherkennung von Schullaufbahngefährdung (vgl. Karagiorgi & Petridou, 2019)
Das von Karagiorgi und Patridou beschriebene Programme for Functional Literacy (PfL) ermöglicht durch formative standardisierte Testungen die Sammlung und Weiterleitung von Leistungsinformationen der Schüler:innen an verschiedene politische und schulische Ebenen. Im unterrichtlichen Sinne können Klassenlehrer diese Daten nutzen, um individuelle Maßnahmen zur Förderung gefährdeter Schüler:innen zu treffen. Auch hier fehlen wieder Informationen zur konkreten Förderung der Schüler:innen auf Grund-lage der Ergebnisse des PfL.

> „[D]espite guidelines to schools about support for students 'at risk', the extent to which these protocols are effective in raising literacy in Cyprus still remains unknown. Hence, light could be shed into the actual school and teacher practices to support students 'at risk', the problems involved, professional development needs, as well as the usefulness and effectiveness of the strategies employed" (Karagiorgi & Petridou, 2019, S. 9).

Eine Schwierigkeit bei standardisierten largescale Tests sei die hohe staatliche Autorität, die dabei ausgeübt wird, verbunden mit der Unterminierung von Lehrkräften in ihrer pädagogischen professionellen Tätigkeit. Folglich erfordert die Umsetzung solcher Programme eine enge Kooperation zwischen politischen Instanzen, Schulen und Lehrkräften. Für die Umsetzung entsprechender Fördermaßnahmen wären diesbezüglich einheitliche Professionalisierungsmaßnahmen von Lehrkräften hilfreich.

Suchergebnisse zur konkreten Nutzung von standardisierten Leistungstests im Unterrichtskontext beziehungsweise der Anwendung solcher Tests durch Lehrkräfte zur Aufbereitung von Lernangeboten blieben aus. Daher wurde die Fragestellung im Verlauf der systematischen Literaturrecherche abstrahiert und auf den erweiterten Schulkontext bezogen. In den betrachteten Studien wurden die Testungen, ob nun in Form eines summativen Leistungstests oder in Form von formativen Assessments, fast ausschließlich computergestützt durchgeführt. Formative Assessment-Werkzeuge (in Form von Lernsoftware, CATs oder DFATs) wurden dafür speziell von IT-Expert:innen und Wissenschaftler:innen entwickelt. Eine diesbezügliche Zusammenarbeit mit Lehrkräften konnte aus den Studien nicht abgelesen werden. Grundsätzlich sind die in den Studien dargelegten Assessment- und Testformate nicht für die Konzeption durch eine Lehrkraft ausgelegt, dafür fehlt es den Lehr-kräften an der notwendigen Professionalität in der Testentwicklung. Jedoch können Lehrkräfte in der Interpretation und Nutzbarmachung statistischer Schüler:innendaten, die aus solchen Testungen hervorgehen, geschult werden. In den Studien wurden keine qualitativen Daten zur Nutzung der Informationen aus den Tests durch Lehrkräfte für ihr Feedback und die weitere Unterrichtsgestaltung betrachtet. Solche konkreten Informationen werden jedoch als besonders hilfreich für ein formatives Assessment gesehen. Häufig wurden summative respektive nationale standardisierte Leistungstests als Vorlage für die Gestaltung formativer Testungen genutzt, da diese bereits jeweiligen staatlich geforderten Bildungs-, Leistungs- und Kompetenzstandards determinieren. Die für den formativen Einsatz entwickelten Tests bilden im Gegensatz dazu auch kleinere inhaltliche Themen-komplexe und Einzelkompetenzen aus den jeweiligen Fächern ab. Hauptsächlich setzen die standardisierten Tests und formativen Assessments geschlossene Aufgabentypen ein (Multiple-Choice, Zuordnungs- und Entscheidungsaufgaben) und nur vereinzelt offene beziehungsweise halboffene Aufgabentypen. Eine computerbasierte und damit standardisierte Auswertung ist nur bei zuerst genannten Aufgabentyp möglich. Die anderen Aufgabentypen erfordern die Korrektur durch eine:n Prüfer:in oder im konkreten Unterrichtskontext durch die Lehrkraft. Die Auswertung wäre dabei nur in dem Maße standardisiert, wie präzise

die Auswertungsrichtlinie konzipiert ist. Die Studienergebnisse lassen darauf schließen, dass eine häufigere Durchführung formativer Onlinetests zu besseren Ergebnissen der Schüler*innen in summativen Tests führt. Formative Assessment-Strategien erwiesen sich mehrheitlich sowohl in sprachlichen als auch in mathematisch-naturwissenschaftlichen Fächern als effektiv, mit Ausnahme der Studie von Faber und Visscher, 2018. Bei dieser Studie gibt es aber keinen konkret feststellbaren Grund, weshalb der Einsatz eines Digitalen Formativen Assessment-Tools (DFAT) im Spracherwerb nicht funktionierte. Grundsätzlich sind insbesondere computerbasierte Testoptionen gut in den Unterricht integrierbar, da sie eine schnelle Durchführung und Auswertung ermöglichen und daher keinen zusätzlichen Zeitaufwand erzeugen. Aus den Studien geht hervor, dass diese Form der Tests nur in enger Kooperation mit Wissenschaftler:innen und Expert:innen möglich sind, wobei bisher ein direkter Austausch mit den Lehrkräften über die konkrete Nutzung der Testdaten fehlt. Für solche Untersuchungen müssten prospektiv Lehrkräfte besser geschult und in den Entwicklungsprozess eingebunden werden. Außerdem müsse dafür ein passender Organisationsrahmen gegeben sein. Darunter fallen beispielsweise die zeitliche Komponente sowie die technische Ausstattung der Schulen.

5 Diskussion

Ausgehend von der Fragestellung, wie standardisierte Leistungstests optimal zur Unterstützung vom Lernen im Unterricht durch Lehrkräfte genutzt werden können, gilt das Erkenntnisinteresse dieser systematischen Literaturrecherche der Untersuchung des aktuellen Forschungsstandes und diesbezüglicher Ergebnisse standardisierter Leistungstests zur Nutzung für ein formatives Assessment in der Schule und welche Bedingungen für einen solchen Einsatz der Tests erfüllt sein sollten.

Bei Betrachtung der Studien war auffällig, dass sowohl die standardisierten Leistungstests als auch die formativen Assessment-Angebote computerbasiert durchgeführt wurden, beispielsweise in Form von Computeradaptiven Tests (CATs), Digitalen Formativen Assessment-Tools (DFATs) oder mittels Lernsoftware. In diesem Zusammenhang wird in den Studien dargelegt, dass die Testdurchführung durch die computerbasierte Unterstützung effizienter werden kann. Problematisch ist, dass nicht alle Schulen über die notwendige technische Ausstattung, beispielsweise in Form von Tablets, verfügen. Dem gegenüber steht jedoch der Aspekt der Nachhaltigkeit, da durch digitale Test- und Assessment-Angebote Ressourcen und Druckkosten gespart werden könnten (vgl. Sutter et al., 2020, S. 41; Clemens et al., 2015). Auch die Speicherung und Verwaltung von Daten und Informationen zu den Schüler:innen und deren Leistungen wäre auf digitalem Weg sicherer und effizienter, ist aber auch professionell abzudecken. Außerdem könnten Bewertungsprozesse beschleunigter oder gar automatisiert erfolgen, zumindest wenn es sich um geschlossene Aufgabentypen handelt. Offene Aufgaben-Items dagegen, die häufig in Leistungsüberprüfungen, insbesondere in sprachlichen Fächern, eingesetzt werden, „sind im Allgemeinen schwieriger als [geschlossene] Auswahlitems, weil hier keine Antwortoptionen vorliegen, die Hinweise auf die richtige Antwort geben könnten“ (Böhme & Engelbert & Weirich, 2017, S. 26). Demzufolge könnten computerbasierte Tests für Lehrkräfte eine zeitliche Entlastung darstellen, sodass diese die Möglichkeit hätten, sich intensiver auf die individuelle Förderung der Schüler:innen im Unterricht zu fokussieren (vgl. Böhme et al., 2017, S. 26). Computergestützt durchgeführte formative Assessments ermöglichen es, diagnostische Daten zu kompletten Lernverläufen zusammenzutragen (vgl. Faber & Visscher, 2018, S. 1). „Hilfreich scheinen [...] solche computergestützten Systeme, die den Lehrkräften explizite Empfehlungen zur Interpretation und Nutzung von diagnostischer Information zu Lernverläufen geben“ (Souvignier et al., 2021, S. 135; Connor, 2019). Dafür müssten

Lehrkräfte jedoch besser in der Nutzung und Analyse digitaler Test- und Assessment-Daten ausgebildet werden, „to ensure a successful pathway to personalized instruction by using the assessment data as tools for screening, benchmarking, and decision making“ (Sutter et al. 2020, S. 49). Weiterhin ermöglichen es Computeradaptive Tests, Lernprozesse differenzierter, angepasst an die individuelle Lerngeschwindigkeit und die Lerntypen der Schüler:innen, zu begleiten. Außerdem können CATs bereits das jeweilige Leistungsniveau einzelner Schüler:innen identifizieren (vgl. Sutter et al., 2020; Mioduser et al., 2001). „Thus, it is crucial to make use of the assessment data not solely for achievement screening and benchmarking, but also invest efforts in the subsequent personalized instruction“ (Sutter, et al., 2020, S. 40 f.).

Auf Grundlage dieser diagnostischen Informationen, könnten Lehrkräfte den Lernprozess im Unterricht gezielter und individualisierter steuern (vgl. Sutter et al. 2020; Mioduser et al., 2001). Wie die Lehrkräfte im Kontext der Studien ihren Unterricht aufgrund der formativen Interventionen gestalteten und welche Informationen sie verwendeten, bzw. ob sie die Ergebnisse aus den Tests und Assessments überhaupt nutzten, wird in den untersuchten Studien nicht in Betracht gezogen oder reflektiert. Zur Optimierung von computergestützten formativen Tests und Assessments sowie der Unterrichtsgestaltung durch die Lehrkräfte und folglich zur Verbesserung von Lernprozessen ist dieser Aspekt jedoch unabdingbar. In der zukünftigen Forschung sollte die Nutzung von und Reaktion auf diagnostische Information von Leistungsmessungen durch das Lehrpersonal mit untersucht werden. Die eingeschlossenen Studien bestätigen eine bereits von Jürgens (2022) getätigte These, dass die Beschäftigung mit computergestützter Leistungsbeurteilung und Leistungs-messung im Kontext der Schulpraxis zunehmend an Bedeutung gewinnt, um praktischen Problemen bei der Umsetzung von Leistungsmessungen im Unterricht zu entgegnen (vgl. Jürgens, 2022, S. 561). Denn Lehrkräften fehlt es beispielsweise noch an diagnostischem Grundlagenwissen, um qualitativ hochwertige Leistungsbeurteilungen durchzuführen:

> „Für die Meisterung solcher anspruchsvollen diagnostischen Aufgaben benötigen Lehrkräfte aber ausreichende Lerngelegenheiten, in denen sie sowohl fundiertes theoretisches Wissen zur pädagogisch-psychologischen Diagnostik erwerben, als auch erste Fertigkeiten und Fähigkeiten im Prozess des praktischen Diagnostizierens im Kontext von Schule anbahnen können“ (Hesse & Latzko, 2017, S. 10).

Im Prozess der formativen Leistungsbeurteilung ist Feedback eine Kernkomponente, um den Lernprozess gezielt zu lenken und Lehr-Lern-Aktivitäten entsprechend anzupassen (vgl. Black & Wiliam, 1998, S. 7 f.). In den computerisierten formativen Assessments der betrachteten Studien wurde Feedback gezielt eingesetzt. Eine schnelle Rückmeldung, direkt nach der Aufgabenlösung, bestenfalls mit einer Erklärung der korrekten Lösung, erweist sich als besonders wirkungsvoll, da die Schüler:innen hierbei die Möglichkeit haben, direkt durch das Assessment zu lernen (vgl. Kulik & Kulik, 1988; Marzano et al., 2001, S. 96 ff.; Edmentum Research, 2018a, S. 3). Automatisierte Tests können solche Rückmeldungen effizient ermöglichen. Allerdings ist zu beachten, dass auch die individuellen Charakteristika der Feedback-Empfängerin oder des Feedback-Empfängers beeinflussen, inwieweit von der Rückmeldung gelernt wird (vgl. Faber & Visscher, 2018, S. 2). Dazu zählen insbesondere bisherige Kenntnisse und Fähigkeiten, demografische und sozioökonomische Hintergründe (vgl. Zheng et al. 2019) oder auch die Motivation (vgl. Timmers et al., 2013).

Nur in einer Studie (vgl. Karagiorgi & Petridou, 2019) wurde ein standardisierter Leistungstest konkret formativ eingesetzt (vgl. Kapitel 4.2.1, Kategorie 3). Ansonsten dienten diese oftmals staatlichen Tests vielmehr als Vorlage zur Eichung von Interventionen durch formatives Assessment, da solche Tests bereits an den jeweiligen

nationalen Bildungsstandards ausgerichtet sind (vgl. Kapitel 4.1.2, Kategorie 2). Die Nutzung formativer Assessments zur konkreten Vorbereitung auf summative standardisierte Tests (vgl. Kapitel 4.1.2, Kategorie 1) ist bezüglich des Rechercheinteresses weniger relevant, zumal das Absolvieren jährlicher nationaler Jahresend-Leistungstests, wie es in den USA vorherrschend durchgeführt wird, im Kontext des deutschen Bildungssystems nicht in dieser Form üblich ist. Hierbei zeigt sich eine Limitation der Arbeit, da die Ergebnisse des Reviews aus vier verschiedenen Ländern stammen und auf unterschiedlichen bildungspolitischen Voraussetzungen und Bildungssystemen beruhen.

Zheng et al. (2019) verfolgten in ihrer Studie sogar das Ziel, standardisierte Leistungstests komplett durch im Unterricht inkludiertes formatives Assessment zu ersetzen. An dieser Stelle stellt sich die Frage nach den Grenzen von standardisierten Leistungstests im Kontext von Unterricht und Lernförderung. Es ist hierbei erwähnenswert, dass standardisierte Leistungstests nicht die einzige Option zur Überprüfung der Leistungsentwicklung von Schülerinnen und Schülern darstellen. Formatives Assessment kann auch unabhängig von standardisierten Leistungstests erfolgen, denn per Definition beschreibt es nur einen Modus der Nutzung diagnostischer Informationen, die man ebenso mithilfe anderer Methoden und Instrumente neben dem klassischen Leistungstest beziehen kann (vgl. Schmidt, 2020, S. 9, S. 21). Black und Wiliam definieren demnach alle Handlungen als formatives Assessment, in die sowohl Lehrkräfte als auch Schüler:innen im Unterrichtsgeschehen involviert sind und die der Erlangung von Informationen für Feedback dienen, welches wiederum von der Lehrkraft für die Modifikation von Lehr-Lern-Aktivitäten genutzt wird (vgl. Black & Wiliam, 1998). Gemäß dieser Begriffserläuterung haben Black und Wiliam fünf Schlüsselstrategien formativen Assessments zusammengefasst: *1) Transparenz von Lernzielen und Erfolgskriterien; 2) Organisation effektiver Unterrichtsgespräche, -aktivitäten und -aufgaben; 3) lernförderliches Feedback; 4) Aktivierung der Schüler:innen als instruktionale Ressourcen füreinander und 5) Aktivierung von Schüler:innen zur Übernahme von Verantwortung für ihren Lernfortschritt* (vgl. Schmidt, 2020, S. 26). Demgemäß ist davon auszugehen, dass zukünftige Forschung sich vermehrt der Erprobung von unterschiedlichen Assessment-Methoden im direkten Unterrichtskontext widmen wird, weniger der konkreten Nutzung von standardisierten Leistungstests durch Lehrkräfte für Ihre Unterrichtsgestaltung. Diese Tendenz wird bereits durch die Forschungsschwerpunkte und -inhalte der eingeschlossenen Studien dargestellt.

6 Literatur

Alexander, B.; Owen, S. & Thames, C. B. (2020). Exploring Differences and Relationships between Online Formative and Summative Assessments in Mississippi Career and Technical Education. *Asian Association of Open Universities Journal, 15(3),* 335–349. https://doi.org/10.1108/AAOUJ-06-2020-0037 [05.02.2023]

Alexander, P. A. (2020). Methodological Guidance Paper: The Art and Science of Quality Systematic Reviews. *Review of Educational Research, 90(1),* 6–23. https://doi.org/10.3102/0034654319854352 [18.01.2023]

Bayer, M. (2015). Bildung, Leistung und Kompetenz. In A. Lange; C. Steiner; S. Schutter, & H. Reiter (Hrsg.), *Handbuch Kindheits- und Jugendsoziologie* (S. 1–13). Wiesbaden: Springer.

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5(1),* 7–74. https://doi.org/10.1080/0969595980050102 [17.12.2022]

Black, P. & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25(6),* 551–575. https://doi.org/10.1080/0969594X.2018.1441807 [12.01.2023]

Böhme, K.; Engelbert, M. & Weirich, S. (2017). Beschreibung der im Fach Deutsch untersuchten Kompetenzen. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016: Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 20-30). Münster: Waxmann.

Clemens, N. H.; Hagan-Burke, S.; Luo, W.; Cerda, C.; Blakely, A.; Frosch, J.; Gamez-Patience, B. & Jones, M. (2015). The Predictive Validity of a Computer-Adaptive Assessment of Kindergarten and First-Grade Reading Skills. *School Psychology Review, 44(1),* 76–97. https://doi.org/10.17105/SPR44-1.76-97 [16.03.2023]

Connor, C. M. (2019). Using Technology and Assessment to Personalize Instruction: Preventing Reading Problems. *Prevention Science, 20(1),* 89–99. https://doi.org/10.1007/s11121-017-0842-9 [16.03.2023]

Döring, N. (2015). Qualitätskriterien für quantitative empirische Studien. In D. Meister, F. von Gross & U. Sander (Hrsg.). *Enzyklopädie Erziehungswissenschaft Online EEO/ Abschnitt: Methoden der empirischen erziehungswissenschaftlichen Forschung*. Weinheim und Basel: Beltz Juventa. http://www.nicola-doering.de/wp-content/uploads/2015/01/D%C3%B6ring-2015-Qualit%C3%A4tskriterien-f%C3%Bcr-quantitative-empirische-Studien.pdf [08.03.2023]

Edmentum Research. (2018a). Research Report: Impacts of the Use of Study Island Practice and Benchmarks. *Online Submission. Edmentum Research*. https://eric.ed.gov/?id=ED604332 [04.01.2023]

Edmentum Research. (2018b). Research Report: Impacts of the Use of Study Island Practice and Benchmarks -- Reading School District, Pennsylvania. *Online Submission. Edmentum Research.* https://eric.ed.gov/?id=ED604315 [04.01.2023]

Faber, J. M.; Luyten, H. & Visscher, A. J. (2016). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education, 106,* 83–96. https://doi.org/10.1016/j.compedu.2016.12.001 [18.02.2023]

Faber, J. M. & Visscher, A. J. (2018). The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment. *Computers & Education, 122,* 1–8. https://doi.org/10.1016/j.compedu.2018.03.008 [12.01.2023]

Frey, S. & Hartig, J. (2022). Kompetenzdiagnostik. In M. Harring, C. Rohlfs, & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S. 928-937). Münster: Waxmann.

Fürstenau, S. & Gomolla, M. (2012). *Migration und schulischer Wandel: Leistungsbeurteilung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gold, A.; Gawrilow, C. & Hasselhorn, M. (2021). Grundlagen schulpsychologischer Diagnostik. In K. Seifried, S. Drewes, & M. Hasselhorn (Hrsg.), *Handbuch Schulpsychologie: Psychologie für die Schule* (S. 109-117). Stuttgart: Verlag W. Kohlhammer.

Havnes, A.; Smith, K.; Dysthe, O. & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation, 38(1),* 21–27. https://doi.org/10.1016/j.stueduc.2012.04.001 [16.12.2022]

Heil, E. A. (2020). *Methode der Systematischen Literaturrecherche für Haus- und Abschlussarbeiten*. Justus-Liebig-Universität Giessen. https://www.uni-giessen.de/de/fbz/fb09/institute/VKE/nutr-ecol/lehre/SystematischeLiteraturrecherche.pdf [25.10.2022]

Hellrung, K. & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review, 9,* 174–190. https://doi.org/10.1016/j.edurev.2012.09.001 [18.02.2023]

Hesse, I., & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3. Aufl.). Opladen: Verlag Barbara Budrich. http://wwwdb.dbod.de/login?url=https://elibrary.utb.de/doi/book/10.36198/9783838547510 [17.03.2023]

Hofmann, J. (2013). *Erziehungswissenschaften*. Berlin: De Gruyter Saur.

Jürgens, E. (2022). Leistungsbeurteilung. In M. Harring, C. Rohlfs, & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S. 552-562). Münster: Waxmann.

Karagiorgi, Y. & Petridou, A. (2019). National literacy assessment: The identification of students 'at risk' in Cyprus. *Research Papers in Education, 34(1),* 1–13. https://doi.org/10.1080/02671522.2017.1329341 [12.01.2023]

Kingston, N. & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement: Issues and Practice, 30(4),* 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x [12.02.2023]

Kulik, J. A. & Kulik, C.-L. C. (1988). Timing of Feedback and Verbal Learning. *Review of Educational Research, 58(1),* 79–97. https://doi.org/10.3102/00346543058001079 [16.03.2023]

Manyak, P. C. & Manyak, A.-M. (2021). Multifaceted Vocabulary Instruction in a Third-Grade Class: Findings from a Three-Year Formative Experiment. *Reading Psychology, 42(2),* 73–110. https://doi.org/10.1080/02702711.2021.1878678 [04.01.2023]

Marzano, R. J.; Pickering, D. & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement* (1. Aufl.). Alexandria, Virginia USA: Association for Supervision and Curriculum Development.

Mioduser, D.; Tur-Kaspa, H. & Leitner, I. (2001). The learning value of computer-based instruction of early reading skills. *Journal of Computer Assisted Learning, 16(1),* 54–63. https://doi.org/10.1046/j.1365-2729.2000.00115.x [16.03.2023]

Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D. G. & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6(7),* 1-6. https://doi.org/10.1371/journal.pmed.1000097 [14.02.2023]

Page, M. J.; McKenzie, J. E.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; Shamseer, L.; Tetzlaff, J. M.; Akl, E. A.; Brennan, S. E.; Chou, R.; Glanville, J.; Grimshaw, J. M.; Hróbjartsson, A.; Lalu, M. M.; Li, T.; Loder, E. W.; Mayo-Wilson, E.; McDonald, S. & Moher, D. (2021*). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews*. BMJ, n71. https://doi.org/10.1136/bmj.n71 [14.02.2023]

Ponce, H. R.; Mayer, R. E.; Figueroa, V. A. & López, M. J. (2018). Interactive highlighting for just-in-time formative assessment during whole-class instruction: Effects on vocabulary learning and reading comprehension. *Interactive Learning Environments, 26(1),* 42–60. https://doi.org/10.1080/10494820.2017.1282878 [04.01.2023]

Schmidt, C. A. (2020). *Formatives Assessment in der Grundschule: Konzept, Einschätzungen der Lehrkräfte und Zusammenhänge*. Wiesbaden (Heidelberg): Springer VS. https://doi.org/10.1007/978-3-658-26921-0 [16.12.2023]

Schwerdt, T. (2010). *PISA und die Folgen: Wozu ist die Schule da?* (1. Aufl.). Bad Heilbrunn: Verlag Julius Klinkhardt.

Schwippert, K. & Coy, M. (2008). Leistungsvergleichs- und Schulqualitätsforschung. In W. Helsper, & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (S. 387-422). Wiesbaden: VS Verlag für Sozialwissenschaften.

Souvignier, E.; Förster, N. Forthmann, B. & Zeuch, N. (2021). Lernverlaufsdiagnostik. In K. Seifried, S. Drewes, & M. Hasselhorn (Hrsg.*), Handbuch Schulpsychologie: Psychologie für die Schule* (S. 129-137). Stuttgart: Verlag W. Kohlhammer.

Sturm, T. (2016). *Lehrbuch Heterogenität in der Schule* (2. Aufl.). München, Basel: Ernst Reinhardt Verlag. http://wwwdb.dbod.de/login?url=https://elibrary.utb.de/doi/book/10.36198/9783838546155 [12.03.2023]

Sutter, C. C.; Campbell, L. O. & Lambie, G. W. (2020). Predicting Second-Grade Students' Yearly Standardized Reading Achievement Using a Computer-Adaptive Assessment. *Computers in the Schools, 37(1),* 40–54. https://doi.org/10.1080/07380569.2020.1720611 [12.01.2023]

Timmers, C. F.; Braber-van den Broek, J. & van den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & Education, 60(1),* 25–31. https://doi.org/10.1016/j.compedu.2012.07.007 [18.03.2023]

Willems, A. S. (2020). *Leitfaden: Das sytematische Review*. IfE Göttingen. https://www.uni-goettingen.de/de/document/download/86db89757b07af3aee23b6b579d262c8.pdf/Leitfaden_Systematisches%20Review_20220510.pdf [30.10.2022]

Zheng, G.; Fancsali, S. E.; Ritter, S. & Berman, S. R. (2019). Using Instruction-Embedded Formative Assessment to Predict State Summative Test Scores and Achievement Levels in Mathematics. *Journal of Learning Analytics, 6(2),* 153–174. https://doi.org/10.18608/jla.2019.62.11 [04.01.2023]

Zitationsvorschlag:
Zenker, H. D. Systematische Literaturrecherche zur Nutzung von standardisierten Leistungstests im Unterricht. *Schulpraxis Entwickeln – Journal für Forschungsbasierte Schulentwicklung*, *2*(1), 17_43. https://doi.org/10.58652/spe.2023.2.p17_43

Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des vom Sächsischen Landtag beschlossenen Haushaltes.

SACHSEN