Technische Universität Dortmund
Fakultät Erziehungswissenschaft, Psychologie und Bildungsforschung

Kognitive und motivationale Effekte digitaler Medien in Lern- und Leistungssituationen

Kumulative Dissertation zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Vorgelegt von
Thomas Ulrich Brüggemann, M. Sc.
Geboren am 06.11.1992 in Düsseldorf

Erstgutachterin: Prof. Dr. Nele McElvany
Zweitgutachterin: Prof. Dr. Charlotte Dignath

Dortmund, 23.02.2023

Danksagung

Ich möchte meinen aufrichtigen Dank an alle Personen aussprechen, die mich während meiner Promotionszeit unterstützt und begleitet haben. Zunächst gebührt ein großer Dank meiner Erstgutachterin Frau Prof. Dr. Nele McElvany für die Aufnahme in die Arbeitsgruppe, die unterstüzenden Diskussionen und die konstruktiven Rückmeldungen.

Ein herzliches Dankeschön gilt auch Frau Prof. Dr. Charlotte Dignath, die als Zweitgutachterin meine Dissertation begutachtet und wertvolle Einsichten beigetragen hat. Einen ganz besonderen Dank widme ich meinen Betreuenden auf Post-Doc Ebene, Frau PD Dr. Ramona Lorenz und Herrn Dr. Ulrich Ludewig. Ihre zahlreichen Rückmeldungen, hilfreiche Hinweise, fruchtbare Diskussionen und die Zusammenarbeit mit ihnen waren während des Promotionsprozesses von unschätzbarem Wert.

Weiterhin bedanke ich mich bei allen Kollegen und Kolleginnen aus der Arbeitsgruppe und dem Institut für die einladende Arbeitsatmosphäre, die anregenden Gespräche, die interessanten Forschungsideen, die gemeinsamme Erschließung von Organisatorischem und die allgemeine Unterstützung.

Zusätzlich möchte ich mich bei den studentischen Hilfskräften und Praktikantinnen für die tatkräftige Unterstützung bei den Datenerhebungen bedanken, ohne die diese Dissertation nicht möglich gewesen wäre.

Zuletzt möchte ich meinen Eltern meinen Dank für ihren Zuspruch aussprechen und meinem Freundeskreis für ihren starken Rückhalt und die wertvolle Programmiernachhilfe.

Ebenso möchte ich Mia und Theia für die Beruhigung und unkonventionellen Perspektiven danken, die sie mir als Katzen geboten haben.

Schließlich danke ich meiner Freundin Laura für die aktive Unterstützung in allen Belangen rund um die Dissertation.

Inhaltsverzeichnis

Zusammenfassung

Digitale Medien halten verstärkt Einzug in das deutsche Schulsystem (Eickelmann et al., 2019; Lorenz, Yotyodying, et al., 2022). Besonders im Rahmen der COVID-19-Pandemie, durch die Schulen zwischen 2020 und 2021 für 74 Tage für den Präsenzunterricht geschlossen waren (Freundl et al., 2021), wurde die Relevanz von digitalen Medien zum Einsatz im Unterricht unterstrichen (SWK, 2022). Um den Unterricht trotz Schulschließungen fortzuführen, musste der Unterricht in Distanz oftmals digital stattfinden (Helm et al., 2021). So konnte der Unterricht mit Wechselunterricht, Konferenzprogrammen, Lernplattformen oder *flipped classrooms* (Akçayır & Akçayır, 2018), bei denen das Lernen außerhalb der Unterrichtsstunden stattfindet, weiterhin umgesetzt werden (Helm et al., 2021). Dabei stellt der Einsatz von digitalen Medien im Unterricht zwar Herausforderungen an die schulischen Ausstattungen und die Kompetenzen der Lehrkräfte, birgt jedoch auch Potentiale in der Gestaltung des Unterrichts (Jude et al., 2020; Lorenz, Brüggemann, et al., 2022). Vor allem im Bereich der Kompetenzerfassung können computerbasierte und computeradaptive Tests standardisierter (Alruwais et al., 2018) und im Falle von adaptiven Tests effizienter (Flens et al., 2016) sein, als analoge Tests auf Papier. Es ist jedoch noch nicht eindeutig geklärt, wie sich digitale Medien auf die Einstellungen von Lehrkräften im Unterricht und auf das Testerleben von Schüler*innen in Testsituationen auswirken.

Die vorliegende Dissertation befasst sich mit der Digitalisierung an Grundschulen in Deutschland aus verschiedenen Blickwinkeln. Zum einen werden Erfahrungen und Kompetenzen von Lehrkräften zum Umgang mit digitalen Medien in Lernsituationen berücksichtigt, zum anderen Potentiale von digitalen Medien in Leistungssituationen untersucht. Ein besonderer Fokus liegt auf den kognitiven und affektiv-motivationalen Merkmalen von Schüler*innen im Umgang mit digitalen Medien im schulischen Kontext in Deutschland. In dieser Dissertation werden verschiedene theoretische Überlegungen und Modelle angewandt, die sich zum einen mit dem technologischen Fortschritt im schulischen Kontext befassen, zum anderen kognitive und affektiv-motivationale Prozesse beschreiben. Die Theorien bewegen sich auf der Unterrichtsebene, auf der Ebene der Lehrkräfte und auf der Ebene der Schüler*innen. Auf der Unterrichtsebene bildet das SAMR-Modell (Puentedura, 2006) einen Bezugsrahmen zur Beschreibung der Implementation digitaler Medien. Auf Lehrkraftebene wird das Modell zum *technological-pedagogical content knowledge* (technologisch-pädagogisches Inhaltswissen, TPACK; Mishra & Koehler, 2006)

herangezogen. Auf der Ebene der kognitiven Effekte auf Schüler*innen wird die *Cognitive Load Theory* (Sweller, 2011) und dessen Erweiterung um die *working memory resource depletion* Hypothese (O. Chen et al., 2018) angewandt. Affektiv-motivationale Merkmale der Schüler*innen werden mit dem *additive model of test anxiety* (additives Modell der Testängstlichkeit; Zohar, 1998) einbezogen.

Diese Dissertation nutzt Daten aus zwei verschiedenen Projekten und den damit verbundenen Studien. In Beitrag I wird untersucht, wie Lehrkräfte 2020 und 2021 digitale Medien einsetzten, um Schüler*innen trotz der Schulschließungen weiterhin unterrichten zu können. Lehrkräfte beantworteten zwei querschnittliche Online-Umfragen während der Schulschließungen in 2020 ($N = 2810$) und 2021 ($N = 1774$). In den Umfragen wurde neben allgemeinen demographischen Daten vor allem Kommunikationswege der Lehrkräfte mit ihren Schüler*innen, Einstellungen und Erfahrungen mit digitalen Geräten und das TPACK (Mishra & Koehler, 2006) erfragt. Die Studie untersuchte, wie sich die Einstellungen und das Fachwissen der Lehrkräfte mit digitalen Medien im Unterricht zwischen den Schulschließungen wandelte und welche Veränderungen es im Einsatz von Methoden im Distanzunterricht gab. Es wurde gezeigt, dass Lehrkräfte zur Kommunikation mit Schüler*innen 2021 deutlich öfter virtuelle Treffen und Lernplattformen nutzten, als es 2020 der Fall war. Mehr als drei Viertel der Lehrkräfte führten 2021 Unterrichtsstunden digital durch, während dieser Anteil in 2020 noch unter einem Drittel der Lehrkräfte lag. Weiterhin schätzten sich Lehrkräfte bezogen auf ihr TPACK in 2021 kompetenter ein als in 2020. Die Studie zeigte, wie sich der Unterricht durch die Schulschließungen auf digitaler Ebene zwischen 2020 und 2021 wandelte. Weiterhin stellte die Studie dar, dass im Rahmen des TPACK Modells die aktive Nutzungvon digitalen Medien im Unterricht zu einer Zunahme des wahrgenommenen TPACK von Lehrkräften führen kann.

Bezogen auf Leistungssituationen im Unterricht mit digitalen Medien wird in Beitrag II untersucht, wie sich digitale Testformate auf Grundschüler*innen in Leistungssituationen auswirken. Verglichen wurden sowohl papierbasierte Tests (PPT), computerbasierte Tests (CBT) als auch computeradaptive Tests (CAT) miteinander. Untersucht wurde spezifisch, inwiefern sich die kognitive Belastung auf Basis der *cognitive load theory* (Sweller, 2011) zwischen den Testformaten unterscheidet. Die Studie nutzte einen Datensatz[1] des Projektes

[1] Für Beitrag 1 wurden Daten aus dem Projekt LL-digital genutzt, die bis zum Stichtag des 31.12.2021 erhoben werden konnten.

Digitale Medien in Lern und Leistungssituationen – Lesekompetenz und Wortschatz im Fokus (*LL-digital*; *N* = 212), in dem die Lesekompetenz von Schüler*innen aus vierten Klassen in allgemeinbildenden Grundschulen in Nordrhein-Westfalen gemessen wurde. Schüler*innen wurden zufällig einem von drei Testformaten (PPT, CBT, oder CAT) zugewiesen. Vor dem Test wurden demographische Angaben erfasst, in der Mitte und am Ende des Tests die kognitive Belastung. Ein linear mixed-effect model zeigte, dass die kognitive Belastung in allen Testformaten über den Testverlauf hinweg anstieg. Dabei konnten keine Unterschiede zwischen den Testformaten festgestellt werden. Allerdings gab es statistisch gestützte Hinweise für die Annahme, dass die kognitive Belastung stärker in adaptiven Tests ansteigt als in nicht-adaptiven Tests (fixed-item tests, FIT). Da CATs weniger Aufgaben stellen müssen, um die gleiche Messpräzision eines FITs zu erreichen (Davey, 2011), sollte die kognitive Belastung von Schüler*innen am Ende eines CATs jedoch nicht höher sein, als in einem FIT mit gleicher Messpräzision. Die Ergebnisse legen weiterhin nahe, dass CATs für alle Fähigkeitsniveaus ähnlich anspruchsvolle Aufgaben stellen, sodass leistungsschwache Schüler*innen im CAT eine geringere kognitive Belastung verspüren als im FIT, während leisungsstarke Schüler*innen eine höhere kognitive Belastung im CAT verspüren, als im FIT.

Beitrag III erweitert die in Beitrag II angesprochenen Unterschiede zwischen PPTs, CBTs und CATs im Hinblick auf das Testerleben in Form von Testängstlichkeit und Lesemotivation. Verwendet wurden ebenfalls Daten der Studie LL-digital[2] (*N* = 387). Spezifisch wurde untersucht, inwiefern sich die Testängstlichkeit und die Lesemotivation von Viertklässler*innen in einem Lesekompetenztest über den Testverlauf entwickelt und ob es Unterschiede zwischen papierbasierten, computerbasierten oder computeradaptiven Tests gibt. Zu Grunde lag das additive Modell der Testängstlichkeit nach Zohar (1998), in dem Testängstlichkeit in eine Eigenschaftskomponente (Trait) und eine Zustandskomponente (State) aufgeteilt wird. Auch für die Lesemotivation wurde diese Aufteilung in eine Zustands- und Eigenschaftskomponente angenommen (Helm & Warwas, 2018; Penk et al., 2014). Daher wurden die dispositionale Testängstlichkeit und die Lesemotivation (Traits) vor dem Test als Eigenschaftsvariablen erfasst, die Zustandsangst und situative Lesemotivation (States) in der Mitte und am Ende des Lesekompetenztestes. Die Zustandsangst wurde zusätzlich direkt vor dem Test erhoben. Eine Ko-Varianzanalyse mit Messwiederholung für Zustandsangst unter Kontrolle von Testängstlichkeit als Eigenschaft und Zustandsangst vor dem Test ergab keine

[2] Für den zweiten Beitrag wurden alle Daten des Projektes LL-digital zum Stichtag des 31.12.2022 genutzt.

Unterschiede zwischen den Testformaten oder Messzeitpunkten. Eine ähnliche Analyse mit Lesemotivation als Eigenschaft und situativer Lesemotivation (State) in der Mitte und am Ende des Tests als Variablen mit Messwiederholungen ergab hingegen, dass Schüler*innen, die am Computer getestet wurden (d.h. CBT und CAT) eine höhere Lesemotivation während des Testes hatten als die Schüler*innen, die am Papier getestet wurden. Diese erhöhte Motivation am Computer nahm jedoch über den Testverlauf ab und glich sich am Ende des Tests an die Motivation des Lesetests am Papier an.

In der studienübergreifenden Diskussion wird darauf eingegangen, wie Lehrkräfte im Kontext des Distanzunterrichtes und Schüler*innen in Leistungssituationen digitale Medien erleben. Die Dissertation stellt dar, dass die Notwendigkeit, digitale Geräte im Unterricht einzusetzen, zu einer erhöhten Nutzung von digitalen Medien für Unterrichtszwecke im zweiten Jahr der Schulschließungen führte und dass Lehrkräfte sich bezüglich ihres Wissens zum Einsatz digitaler Medien (in Anlehnung an das TPACK-Modell) als kompetenter einschätzten. Weiterhin wird gezeigt, dass sich digital gestützte Testformate im Grundschulunterricht erfolgreich einsetzen lassen und dass computerbasierte und computeradaptive Tests bei der Kompetenzerfassung der Lesefähigkeit keine negativen Auswirkungen in Form von Testängstlichkeit oder kognitiver Belastung haben, sondern stattdessen die Motivation von Grundschüler*innen (kurzfristig) positiv beeinflussen können.

1. Einleitung

Digitale Medien bieten im schulischen Kontext sowohl für den Unterricht, als auch für die Diagnostik und Kompetenzmessung besondere Potentiale und Chancen (Alruwais et al., 2018; Lorenz, Brüggemann, et al., 2022). Ein verstärkter Einsatz von digitalen Medien an Schulen in Deutschland ist ein erklärtes Ziel der Bildungspolitik (KMK, 2016, 2021). Die Notwendigkeit des Ausbaus der digitalen Infrastruktur wurde aufgrund der COVID-19-Pandemie, in deren Zuge Schulen in Deutschland zwischen März 2020 und Mai 2021 insgesamt 74 Tage lang bundesweit vollständig geschlossen waren (Freundl et al., 2021), auch von bildungspolitischer Seite erneut bekräftigt (SWK, 2022). Um während dieser Zeit die Qualifikation von Schüler*innen als gesellschaftliche Funktion der Schule weiter wahrnehmen zu können (Fend, 2009) bedurfte es dem Einsatz digitaler Medien. Sie wurden genutzt, um mit Schüler*innen weiterhin kommunizieren zu können, Aufgaben zu erteilen und Unterrichtsstunden durchführen zu können (Helm et al., 2021). Neben der Qualifikation von Schüler*innen ist die Kompetenzerfassung ein zentrales Aufgabenfeld der Schule. Kompetenztests können sowohl zur Diagnostik im Rahmen der Qualifikationsfunktion genutzt werden, als auch um Schüler*innen gemäß der Selektions- und Allokationsfunktion der Schule zuweisen zu können (Fend, 2009). Digitale Medien können auf verschiedene Möglichkeiten zur Kompetenzerfassung eingesetzt werden: Computerbasierte Testformate können Testpersonen automatisch bewerten (Alruwais et al., 2018) und computeradaptive Testformate können Tests effizienter und kürzer gestalten als papierbasierte Tests (Davey, 2011; Embretson & Reise, 2009). Im Bereich Lesen wirken digitale Medien oft motivierend, sowohl im Unterricht (Lorenz, Brüggemann, et al., 2022), in Testsituationen (Chua, 2012) oder zuhause (Picton, 2014). Dennoch werden digitale Medien an Schulen in Deutschland im Unterricht selten zur Kompetenzerfassung eingesetzt (Fraillon et al., 2019). Weiterhin stellen digitale Medien Lehrkräfte auch vor Herausforderungen. Neben der Voraussetzung an die Fähigkeit, mit digitalen Medien sowie unterrichtsrelevanten Programmen oder Apps umgehen zu können, ist vor allem die Verlässlichkeit der Technik ein Grund zur Ablehnung (Lorenz, Brüggemann, et al., 2022). Besonders im Bereich der Lesekompetenz zeigten Studien, dass Leser*innen, die an einem Bildschirm lesen, geringere Leistungen in darauffolgenden Leseverständnistests aufweisen als Leser*innen, die am Papier lesen (Delgado et al., 2018; Furenes et al., 2021). Zudem findet sich in der Literatur Kritik an den Auswirkungen von adaptiven Tests auf das Testerleben (Colwell, 2013). Testeigenschaften, die bei adaptiven

Tests notwendig sind, wie zum Beispiel eine fixierte Lösungsrate oder das Fehlen der Möglichkeit zu vorherigen Aufgaben im Test zurückzugehen, können sich negativ auf das Testerleben auswirken (Ling et al., 2017; Ortner & Caspers, 2011).

Vor diesem Hintergrund beschäftigt sich die vorliegende Dissertation mit der Frage, wie sich digitale Medien auf die motivationalen und kognitiven Aspekte von Lehrkräften und Schüler*innen im Unterricht auswirken. Dabei wird untersucht, wie sich die Einstellungen und Fähigkeitsselbstkonzepte der Lehrkräfte in Deutschland in Zeiten des Distanzunterrichts zwischen 2020 und 2021 im Hinblick auf den Einsatz digitaler Medien im Unterricht wandelten (Beitrag I). Zudem wird erforscht, inwiefern sich computerbasierte und computeradaptive Testformate im Unterricht eignen und welche Vorteile oder Nachteile sie in Hinblick auf das Testerleben mit sich bringen (Beiträge II und III).

Diese Dissertation beginnt mit dem theoretischen Hintergrund (Kapitel 2), in dem spezifisch auf den Einsatz digitaler Medien in Lernsituationen (Kapitel 2.1) und in Leistungssituationen (Kapitel 2.2) eingegangen wird. Dabei liegt der Fokus zum einen auf dem allgemeinen Einsatz von digitalen Medien im Unterricht (Kapitel 2.1.1) und besonders auf dem Einsatz von digitalen Medien während der COVID-19-Pandemie (Kapitel 2.1.2). Unterschiede zwischen verschiedenen Testmedien, wie papierbasierte, computerbasierte und computeradaptive Tests, werden in Kapitel 2.2.1 dargestellt. Die Einflüsse, die diese Testformate auf das Testerleben haben können, werden im darauffolgenden Kapitel 2.2.2 beschrieben. An den theoretischen Hintergrund knüpft die Darstellung der Forschungsfragen an, die sich für diese Dissertation aus der Theorie ergeben (Kapitel 2.3). Der Theorie folgen die Einzelbeiträge (Kapitel 4). In der Gesamtdiskussion (Kapitel 5) werden die empirischen Beiträge zunächst zusammengefasst (Kapitel 5.1) und dann beitragsübergreifend diskutiert (Kapitel 5.2). Die Stärken und Limitationen der Dissertation werden im darauffolgenden Abschnitt thematisiert (Kapitel 5.3), woraufhin Implikationen für die Forschung und schulische Praxis abgeleitet werden (Kapitel 5.4). Schließlich endet die Dissertation mit einem Fazit (Kapitel 5.5).

2. Theoretischer Hintergrund

In dieser Dissertation werden verschiedene Themenfelder untersucht, in denen sich die Nutzung von digitalen Medien auf die motivationalen und kognitiven Aspekte der Nutzer*innen im Schulkontext auswirken können. Untersucht werden sowohl die Einflüsse auf der Ebene der Lehrkräfte, als auch die Ebene der Schüler*innen in Tests als spezifische

Unterrichtssituation. Dementsprechend baut diese Dissertation auf mehreren Theorien auf, die sich mit den einzelnen Themen beschäftigen. Auf der Ebene der Lehrkräfte in Lernsituationen wird vor allem die Selbstwirksamkeitserwartung zum Einsatz von digitalen Medien im Unterricht im Rahmen des TPACK-Modells betrachtet (Koehler & Mishra, 2009). Auf der Ebene der Schüler*innen werden sowohl motivationale als auch kognitive Einflüsse digitaler Medien berücksichtigt. Kognitive Einflüsse werden mithilfe der Cognitive Load Theory (Sweller et al., 2019) betrachtet, die einen Ansatz bildet, um Lern- und Arbeitsprozesse zu erklären. In diesem Rahmen wird auch die *working memory ressource depletion* Hypothese genutzt, um Verläufe von kognitiver Belastung über Leistungssituationen hinweg zu erklären. Weiterhin werden für die Effekte auf motivationaler und affektiver Ebene die situativen Erfahrungen unter Berücksichtigung von Eigenschaften und Prädispositionen berücksichtigt. Dies folgt bei der Testängstlichkeit dem additiven Model nach Zohar (1998).

2.1 Digitale Medien in Lernsituationen

Digitale Medien lassen sich im Unterricht vielseitig einsetzen. Möglichkeiten umfassen unter anderem die Recherche, Sprachförderung oder Kompetenzmessung (Lorenz, Brüggemann, et al., 2022). Einen Ansatz zur Klassifizierung, wie digitale Medien in den Unterricht integriert werden können bildet das SAMR Modell (Puentedura, 2006). Dieses unterscheidet zwischen der Substitution, Augmentation, Modifikation und Redefinition von Unterrichtsinhalten mit digitalen Medien. Bei der Substitution werden analoge Methoden und Materialien mit digitalen Medien ersetzt, ohne dass es funktionale Unterschiede gibt. Beispielsweise schreiben Schüler*innen Texte mit einem Textverarbeitungsprogramm am Computer, anstelle auf Papier. Bei der Augmentation werden Möglichkeiten digitaler Medien genutzt, um Unterrichtsaufgaben zu verbessern. So können Schüler*innen die im Textverarbeitungsprogramm selbst geschriebenen Texte selbstständig mithilfe einer Rechtschreibprüfung überprüfen. Zur Modifikation von Unterricht mit digitalen Medien werden Aufgaben so angepasst, dass sie auf digitale Medien zugeschnitten sind. Zum Beispiel kann eine Gruppe von Schüler*innen instruiert sein, einen kollaborativen Text zu produzieren, wodurch kollaborative Schreibprogramme im Unterricht genutzt werden können. Die Redefinition schließlich beschreibt Lernprozesse und Unterrichtsaufgaben, die ohne den Einsatz digitaler Medien nicht möglich wären. So können die von Schüler*innen produzierten Texte hinterher digital veröffentlicht werden, damit andere Personen Rückmeldungen geben können, von denen die Schüler*innen lernen können. Das Modell ist hierarchisch angeordnet,

wobei Methoden, die digitale Medien stärker in den Unterricht integrieren, als inhärent lernfördernder angesehen werden. Dagegen ist eine Umstellung zu digitalen Medien nicht automatisch lernfördernd (Cess et al., 2018). Auch Lehrkräfte geben an, dass eine reine Substituierung von analogen Unterrichtsinhalten mit digitalen Medien keinen inhärenten Mehrwert für den Unterricht mit sich bringt (Lorenz, Brüggemann, et al., 2022). Diese normative Wertung von Methoden, sowie die rigide Struktur wurden in der Vergangenheit kritisiert, da eine Integration von digitalen Medien in den Unterricht entsprechend einer höheren Stufe im Rahmen des SAMR-Modells nicht automatisch mit einer Verbesserung des Unterrichts einhergehen muss (Hamilton et al., 2016). Nichtsdestotrotz lässt sich die Taxonomie des SAMR-Modells anwenden, um Unterschiede im Einsatz von digitalen Medien im Unterricht darzustellen.

Obgleich es viele verschiedene Einsatzmöglichkeiten von digitalen Medien im Unterricht gibt (Jude et al., 2020), wurden sie an deutschen Schulen vor der COVID-19-Pandemie selten eingesetzt (Fraillon et al., 2019). Bedeutsam für den Einsatz von digitalen Medien im Unterricht ist die Motivation der Lehrkräfte, wie im Rahmen des *technological acceptance model* (Davis, 1985; Marangunić & Granić, 2015) dargelegt. Diese Motivation lässt sich in die wahrgenommene Nützlichkeit und die wahrgenommene Benutzerfreundlichkeit einteilen. Beide wirken sich auf die Einstellung digitale Medien zu nutzen aus. In diesem Modell sind somit zum einen die Einstellungen von Lehrkräften gegenüber digitalen Medien (Petko, 2012) und zum anderen deren Selbstwirksamkeit mit digitalen Medien unterrichten zu können als Aspekt der Benutzerfreundlichkeit wichtige Voraussetzungen, um die Nutzung von digitalen Medien im Unterricht vorherzusagen (Paraskeva et al., 2008). Das TPACK-Modell liefert einen Ansatz, diese Selbstwirksamkeit zu erfassen (Mishra & Koehler, 2006). Dabei benötigten Lehrkräfte neben ihrem Fachwissen Kompetenzen im Hinblick auf den Umgang mit digitalen Geräten und sie müssen in der Lage sein, das Fachwissen mit technologischen Mitteln pädagogisch wertvoll in den Unterricht einzubinden (Mishra & Koehler, 2006). Diese Beziehung zwischen technologischem, pädagogischem und Fachwissen ist im TPACK-Model abgebildet (Koehler & Mishra, 2009). Dem Modell zufolge verfügen Lehrkräfte über separate Wissensdomänen im Bereich der Pädagogik, Technik und dem Unterrichtsfach, die insbesondere in ihrer kombinierten Betrachtung wertvoll sind. Pädagogisches Inhaltswissen ist beispielsweise das Verständnis darüber, wie Unterrichtsinhalte Schüler*innen verständlich vermittelt werden können. Fachinhaltswissen ist Wissen über das Unterrichtsfach, das die Lehrkraft unterrichtet.

Technisches Inhaltswissen ist Wissen über den Umgang mit technischen Geräten und digitalen Medien. Vor allem das technische Inhaltswissen befindet sich ständig im Wandel, da sich neue Programme und Methoden sehr schnell entwickeln (Koehler & Mishra, 2009). Für einen gezielten Einsatz von digitalen Medien im Unterricht ist im TPACK-Modell die Schnittstelle der drei Wissensdomänen relevant, da hier Lehrkräfte in der Lage sind ihr Fachwissen mithilfe von digitalen Medien auf pädagogisch wertvolle Art und Weise mitzuteilen.

2.2 Digitale Medien während der COVID-19-Pandemie

Die COVID-19-Pandemie ist eines der einschneidendsten gesellschaftlichen Ereignisse in der modernen Geschichte (Cruz-Cárdenas et al., 2021). Zur Eindämmung der Pandemie wurden in Deutschland weitreichende Maßnahmen zur Beschränkung der Infektionen wie Maskenpflichten, Homeoffice, Ausgangssperren oder Testpflichten entschieden (BMG, 2022). Während Zeiten mit einer hohen Rate an Neuinfektionen pro 100.000 Einwohner*innen pro Woche (7-Tage-Inzidenz) wurden Schulen zwischen März 2020 und Mai 2021 zeitweise geschlossen (Freundl et al., 2021). Unterricht fand in dieser Zeit entweder mit sich abwechselnden Klassenhälften („Wechselunterricht“; MSB NRW, 2022), oder digital statt („Distanzunterricht“; Eickelmann & Gerick, 2020). Besonders in Zeiten der Schulschließungen war der Einsatz von digitalen Medien im Unterricht von besonderer Bedeutung. Unterricht konnte mit Videokonferenzen durchgeführt und Aufgaben sowie Unterrichtsmaterialien konnten mittels Lernplattformen bereitgestellt werden (Jude et al., 2020).

Grundsätzlich waren Schulen in Deutschland auf die Anforderungen der Schulschließungen und dem damit verbundenen Distanzunterricht im internationalen Vergleich nicht gut vorbereitet (Freundl et al., 2021). Schon vor dem Ausbruch der Pandemie lag Deutschland im Bereich der Digitalisierung an Schulen im internationalen Vergleich zurück (Fraillon et al., 2019). So nutzten nur 23 Prozent der in Deutschland befragten Lehrkräfte in der ICILS Studie 2018 digitale Medien zum Unterrichten, deutlich unter dem Mittelwert der Gesamtstichprobe von 48 Prozent. Weiterhin hatten in Deutschland nur 26 Prozent der Achtklässler*innen einen Internetzugang an der Schule, während alle Schüler*innen in Dänemark oder Finnland Internetzugang hatten. Vor diesem Hintergrund ist es nicht überraschend, dass sich die Schulschließungen negativ auf verschiedene Aspekte des schulischen Lehrens und Lernens ausgewirkt haben (Helm et al., 2021). In

Selbsteinschätzungen gaben Schüler*innen an, dass sie weniger Lernen würden (Baier & Kamenowski, 2020), schlechtere Leistungen erbrächten und sich Sorgen über ihre Schulleistungen machten (Trültzsch-Wijnen & Trültzsch-Wijnen, 2020). Dies spiegelte sich auch in Befunden des IFS-Schulpanels wider, in dem Schüler*innen der vierten Klasse 2021 im Durchschnitt geringere Leseleistungen erbrachten, als Schüler*innen 2016 (Ludewig, Kleinkorres, et al., 2022).

Unter Berücksichtigung der Ausgangslage stellt sich die Frage, inwiefern die COVID-19-Pandemie die Digitalisierung an Schulen in Deutschland hat voranschreiten lassen. Da der Einsatz digitaler Medien im Unterricht teilweise unumgänglich wurde, ist zu erwarten, dass Lehrkräfte digitale Medien über die COVID-19-Pandemie verstärkt für Unterrichtszwecke eingesetzt haben. In diesem Zuge könnten sich auch ihre Einstellungen und ihr TPACK verändert haben, was zu einem Anstieg im Einsatz von digitalen Medien im weiteren Unterrichtsverlauf führen kann (Kihoza et al., 2016; Paraskeva et al., 2008).

2.3 Digitale Medien in Leistungssituationen

Leistungssituationen im Schulkontext sind Unterrichtssituationen, in denen Schüler*innen zuvor gelernte Unterrichtsinhalte wiedergeben müssen. Oft finden sie in Form von standardisierten Tests statt, bei denen Schüler*innen Aufgaben in Einzelarbeit bearbeiten müssen, wodurch der Lern- und Leistungsstand erkannt werden kann. Bei dieser Methode der Diagnostik lassen sich digitale Medien wie Computer gezielt einsetzen. Digitale Testformate haben in der schulischen Praxis eine Reihe an Vor- und Nachteilen (Rezaie & Golshan, 2015). Neben den benötigten Endgeräten verlangen computerbasierte Tests von Testadministrator*innen und Testpersonen die Fähigkeit ab, mit dem Computer umgehen zu können. Vor allem in der Vorbereitung von Tests ist technisches Vorwissen eine Voraussetzung für Administrator*innen. Dafür sind computeradministrierte Tests am Testtag einfach zu nutzen, sparen Zeit bei der Auswertung, verringern Kosten und sind ökologisch nachhaltiger, da sie weniger Papier verbrauchen (Alruwais et al., 2018). Weiterhin erlauben Computer die Nutzung von computeradaptiven Tests, bei denen sich die Aufgabenschwierigkeit an die Fähigkeiten der zu testenden Personen anpasst. Diese sind deutlich effizienter als Tests mit einer vorher festgelegten Aufgabenabfolge (Davey, 2011), wodurch die Testlänge reduziert werden kann. Daher werden computerbasierte Tests verstärkt in Large Scale Assessments (Hußmann et al., 2017) eingesetzt. Auch die PISA-Studie 2018 nutzte einen *multistage adaptive test* (MAT) im Bereich Lesen, bei dem Itemblöcke in

Abhängigkeit der Fähigkeit der Person eingesetzt werden (Yamamoto et al., 2019). Dabei ist jedoch zu beachten, dass eine Änderung des Testformates von PPT zu CBT oder CAT für die zu testenden Personen ebenfalls Veränderungen mit sich bringt (Colwell, 2013; Noyes & Garland, 2008). Mehrere Studien untersuchten, inwiefern sich Testleistungen zwischen den Testformaten unterscheiden, mit teils gemischten Ergebnissen (Kong et al., 2018; Wang et al., 2007). Vor allem im Bereich Lesen ließen sich Unterschiede ausmachen. Mehrere Metaanalysen konnten aufzeigen, dass das Lesen am Bildschirm zu einer geringeren Leistung in Lesekompetenztests führt (Clinton, 2019; Delgado et al., 2018). Diese sogenannte *screen inferiority* ließ sich allgemein bei Schüler*innen (Furenes et al., 2021), sowie spezifisch bei Grundschüler*innen finden (Lenhard et al., 2017). Allerdings gibt es noch keinen wissenschaftlichen Konsens zur Ursache dieser *screen inferiority*. Diskutiert werden neben perzeptuellen Gründen (Mayr et al., 2017) auch motivationale (Colwell, 2013) oder kognitive Faktoren (DeStefano & LeFevre, 2007), die das Testerleben beeinflussen, was sich wiederum auf die Testleistung im (Lese-) Test auswirkt.

2.3.1 Testformate

Die zuvor beschriebenen Testformate von papierbasierten Tests, computerbasierten Tests und computeradaptiven Tests lassen sich in zwei Kategorien einteilen. Zum einen unterscheiden sich papierbasierte und computerbasierte Tests hinsichtlich des *Mediums*, auf dem getestet wird (d.h. Papier oder Bildschirm), zum anderen unterscheiden sich Tests hinsichtlich des *Modus* (d.h. fixiert oder adaptiv). Im Folgenden werden die Unterschiede zwischen den Testformaten erläutert, durch die differentielle Effekte aufgrund von Unterschieden im Medium oder Modus plausibel sind.

Testmedium. Das Testen am Papier unterscheidet sich in mehreren Punkten von dem Testen am Computer. Im Lesekontext lässt sich zwischen dem *digitalen Lesen* und *Lesen an digitalen Geräten* unterscheiden. Beim *digitalen Lesen* werden formatspezifische Funktionen, die am Computer möglich sind, in den Leseprozess einbezogen. Beispiele für formatspezifische Funktionen sind multimodale Textdarstellungen, die visuelle, akustische und räumliche Darstellungen verknüpfen (Rowsell & Burke, 2009), Hypertext, bei dem einzelne Wörter auf neue Textabschnitte verweisen (DeStefano & LeFevre, 2007) oder *Adaptable Books*, bei denen Geschichten selbstständig fortgeführt werden (Hauck-Thum, 2018). Das *Lesen an digitalen Geräten*, das in dieser Arbeit im Fokus steht, bezeichnet hingegen das Lesen von Texten an einem Bildschirm, ohne formatspezifische Funktionen einzubeziehen. Für das Lesen an digitalen Geräten identifizierte schon (Gould, 1968)

technische Herausforderungen, die Unterschiede in der Leseerfahrung zwischen Papier und Bildschirm hervorrufen können. Neben Variablen wie Bildschirmhelligkeit, Pixeldichte (Mayr et al., 2017), Bildschirmauflösung (Köpper et al., 2016), Bildschirmwiederholungsrate, Schriftart (Ukonu et al., 2021) und Farbarten sind auch Unterschiede in den Anforderungen zum Umgang mit den Texten zu benannt. Interaktion mit Texten am Bildschirm verlangt von Leser*innen die Fähigkeit mit Eingabeschnittstellen, wie Maus und Tastatur oder Touchpad umgehen zu können (Alruwais et al., 2018) und längere Texte erfordern, dass Leser*innen regelmäßig scrollen.

Diese inhärenten Unterschiede im Testmedium können sich auf verschiedene kognitive und motivationale Aspekte des Lesens auswirken. Voraussetzungen an die Kompetenz von Leser*innen, digitale Geräte zu bedienen, können Personen mit geringer Computerselbstwirksamkeit verunsichern (Saadé & Kira, 2009). *Hypertexte* zwingen Lesende dazu, regelmäßig die Entscheidung zu treffen, ob mehr Informationen benötigt werden, was sich negativ auf die Prozesse des Arbeitsgedächtnisses auswirken kann und somit das Leseverständnis beeinträchtigt (DeStefano & LeFevre, 2007). Scrollen führt zu einem geringeren Engagement mit narrativen Texten (Mangen & Kuiken, 2014) und zu einem schlechteren Überblick über die Textstruktur (Piolat et al., 1997).

Testmodus. Der Einsatz von Computern beim Testen erlaubt den Einsatz von computeradaptiven Tests. Computeradaptive Tests nutzen die Rechenfähigkeit von Computern, um während des Tests die Fähigkeit einer zu testenden Person zu schätzen. In Abhängigkeit der geschätzten Fähigkeit werden dann die Items administriert, die den höchsten Informationsgehalt über das Fähigkeitsniveau, der zu testenden Person liefern (Embretson & Reise, 2009). Somit sind adaptive Tests effizienter als Tests mit fixierten Itemabfolgen (FITs), da Items mit niedrigem Informationsgehalt, also Items, die beispielsweise viel zu schwierig oder viel zu leicht für die Person sind, nicht administriert werden (Davey, 2011). Hierdurch lassen sich Testlängen substanziell reduzieren (Flens et al., 2016).

Adaptive Tests haben einige Voraussetzungen an die Teststruktur. Zur Auswahl der Aufgaben müssen diese während des Tests (zumindest partiell) eindeutig als richtig oder falsch bewertbar sein (Embretson & Reise, 2009). Weiterhin muss das zu messende Konstrukt eindimensional sein, da man von einer korrekten Antwort in einer Aufgabe auf die Antworten in leichteren Aufgaben schließen können muss. Schließlich müssen Itemparameter, wie die Schwierigkeit der einzelnen Items bekannt sein. Daher basieren adaptive Tests zum Großteil

auf der Item Response Theory (IRT; Moosbrugger, 2012; Michel et al., 2018; für CATs basierend auf *machine learning* siehe Zheng et al., 2020).

IRT geht anders als die klassische Testtheorie davon aus, dass sich Einzelitems in mehreren Aspekten, z.B. der Schwierigkeit voneinander unterscheiden können. Im Rahmen der IRT wird angenommen, dass die Wahrscheinlichkeit dafür, dass eine Testperson ein Item löst von mehreren Parametern vorhergesagt wird. Im Rasch- oder 1-parameter-logistic (1PL) Modell ist die Wahrscheinlichkeit einer richtigen Antwort für ein Item I zum einen von der Itemschwierigkeit I_b, zum anderen von der Fähigkeit der beantwortenden Person J θ_j abhängig. Komplexere Modelle, wie das 2PL-Modell berücksichtigen zusätzlich, inwieweit das Item zwischen Personen mit unterschiedlichen Fähigkeitsniveaus mithilfe des Itemparameters I_a diskriminieren kann (Birnbaum, 1968). Im 3PL-Modell wird zusätzlich ein Rateparameter I_c angenommen, der beschreibt, dass Personen mit sehr niedriger Fähigkeit bei Fragen mit Mehrfachauswahl die korrekte Antwort erraten können (Birnbaum, 1968). Gegenübergestellt geht das 4PL-Model davon aus, dass Personen, die ein Item eigentlich korrekt beantworten würden, aus verschiedenen Gründen (z.B. Selbstüberschätzung oder mangelnde technische Fähigkeit) die falsche Antwort auswählen (Culpepper, 2016).

Wenn Itemschwierigkeit und Diskrimination bekannt sind, lässt sich die Lösungswahrscheinlichkeit für eine Person mit einer beliebigen Fähigkeit darstellen. Dabei bildet die erste Ableitung dieser Itemcharakteristikkurven den Informationsgehalt des Items für eine Person. Adaptive Tests schätzen auf Basis der Itemcharakteristiken und dem Antwortverhalten nach jedem beantworteten Item die Fähigkeit der Person und wählen dann ein neues Item aus, das den höchsten Informationsgehalt für eine Person mit dem geschätzten Fähigkeitswert hat (Embretson & Reise, 2009).

Adaptive Tests können sich in ihrer Wirkung auf Testpersonen in mehreren Aspekten von FITs unterscheiden. Um den Informationsgehalt von Items zu maximieren, werden in der Regel Items so ausgewählt, dass Testpersonen eine durchschnittliche Lösungsrate von etwa 50 Prozent über den gesamten Test erhalten (Embretson & Reise, 2009; Ling et al., 2017; Weiss & Betz, 1973). Die durchschnittliche Lösungsrate der Aufgaben über den Test hinweg in einem CAT ist unabhängig von der Leistung, da sowohl starke als auch schwache Testteilnehmende Aufgaben erhalten, die anspruchsvoll genug sind, um einen hohen Informationsgehalt zu haben. In einem FIT hingegen ist die Lösungsrate direkt abhängig von der Fähigkeit, da sich die Aufgaben nicht anpassen können. Eine starke Person wird immer mehr Aufgaben korrekt beantworten als eine schwache Person. Dies kann sich auf das

Testerleben von Testeilnehmer*innen auswirken, vor allem wenn die Selbsteinschätzungen der eigenen Leistung aufgrund von anspruchsvolleren Aufgaben nicht bedient wird (Ortner et al., 2014).

Weiterhin passt sich der Test systematisch an die Fähigkeit der zu testenden Person an, wobei das Fähigkeitsniveau zu Beginn eines adaptiven Tests unbekannt ist (*starting point of entry problem*). Ohne Vorwissen über die Fähigkeiten einer Testperson kann sich der CAT erst nach der ersten Aufgabe an die Fähigkeit der Person anpassen (Embretson & Reise, 2009). Da zum Beginn des Tests weniger Information über die Fähigkeit der Testperson bekannt ist, ist der Standardfehler der Schätzung der Fähigkeit am Anfang größer als zum Ende des Tests. Dadurch unterscheidet sich die Differenz zwischen Fähigkeitsniveau der Testperson und der Schwierigkeit der administrierten Aufgaben systematisch zwischen dem Beginn eines adaptiven Tests und dem Ende eines adaptiven Tests. Für Testpersonen kann der Test so zu Beginn einen anderen Charakter haben, da Items stärker zwischen einfach und schwer schwanken als am Ende des Tests, wenn Items ähnlich anspruchsvoll sind.

CATse. Ein wichtiger Aspekt in Beiträgen II und III ist, dass die Aufgabendarstellungen am Papier und am Computer möglichst ähnlich sind, um Unterschiede zwischen den Testformaten auf die Testformate selbst herleiten zu können und darstellungsbezogene Variablen, wie Schriftart, Schriftgröße, Textrahmen oder Scrolling als Ursachen für Effekte ausschließen zu können. Daher wurde im Zuge dieser Dissertation das Programm CATse (Computer Adaptive Test Structural Environment) entwickelt, dass für die Testadministration des CBT und CAT eingesetzt wurde. Die Eigenerstellung des Programmes in Python 3.7 erlaubte es, die Darstellungen zwischen der Papierversion und den Computerversionen mithilfe der Bibliothek *tkinter* (Lundh, 1999) anzupassen. Der adaptive Test basierte auf der *Fairen und adaptiven Lesekompetenzdiagnose* (FALKE) für die dritte und vierte Klasse (McElvany & Schwabe, 2019; Ludewig, Trendtel, et al., 2021). Der FALKE besteht aus 85 Texten, zu denen es 132 Items gibt. Texte sind sachlich oder narrativ und Aufgaben sind inferenz- oder textbasiert. Die korrekten Antworten in textbasierten Aufgaben stehen explizit im Text, während bei inferenzbasierten Aufgaben anhand des Textes auf die korrekte Antwort geschlossen werden muss. Daher gibt es vier verschiedene Itemformate, die in Tabelle 1 dargestellt sind:

Tabelle 1

Itemformate der FALKE.

	Narrativer Text	Sachtext
Textbasierte Aufgabe	Narrativer Text mit textbasierter Aufgabe	Sachtext mit textbasierter Aufgabe
Inferenzbasierte Aufgabe	Narrativer Text mit inferenzbasierter Aufgabe	Sachtext mit inferenzbasierter Aufgabe

Die FITs umfassten 25 Items, die anhand von mehreren Charakteristiken ausgewählt wurden: Zum einen wurden die Itemformate balanciert, sodass jede der vier Arten an Items in Tabelle 1 etwa gleich oft vorkam. Weiterhin wurden Items für den FIT gewählt, die im CAT eine hohe Wahrscheinlichkeit hatten ausgewählt zu werden. Zuletzt wurden Items gewählt, sodass die Itemschwierigkeiten I_b normalverteilt waren. Für die Itemselektion im CAT wurde die *maximum Fisher information* (PFI) für jedes Item für eine Person *j* mit dem Fähigkeitswert θ_j geschätzt (Barrada et al., 2009). Der Fähigkeitsschätzer (θ) einer Person (j) θ_j im CAT wurde mithilfe der *expected a posteriori* (EAP; Bock & Mislevy, 1989) berechnet. Der EAP wurde als Fähigkeitsschätzer gewählt, da er im Vergleich zu nicht-baysianischen Verfahren wie *maximum likelihood estimates* (MLE) besser mit extremen Antwortverhalten umgehen kann und der EAP im Vergleich zu anderen bayesianischen Verfahren geringere systematische und zufällige Schätzfehler produziert (Wang & Vispoel, 1998).

Als bayesianisches Verfahren können beim EAP Vorannahmen über die Testperson in die Berechnung von θ_j einbezogen werden. Im FALKE bestand eine a priori Annahme zur Lesekompetenz der Schüler*innen, da der FALKE für die dritte und vierte Klasse geeignet ist, im Rahmen der Studien das Programm CATse jedoch nur bei Schüler*innen der vierten Klasse zum Einsatz kam. Diese *prior* berücksichtigte, dass die getesteten Schüler*innen in der vierten Klasse waren. Anhand der *prior* wurde ein Startwert für θ_j bestimmt, mit dem die Itemselektion durch den PFI beginnen konnte. Bei der Itemselektion nach der PFI wurden die Itemformate aus Tabelle 1 berücksichtigt, sodass die Itemformate im CAT ebenfalls ausbalanciert wurden. Zur Berechnung der *prior* wurde die Gauß-Hermite Integration (Ehrich, 2002) genutzt, die mithilfe der Bibliothek *NumPy* (Harris et al., 2020) berechnet wurde. Weiterhin wurde die Bibliothek *pandas* (McKinney, 2010) zur Organisation der Dataframes genutzt. Zusammenfassend entstand dadurch ein Programm, mit dem die Lesekompetenz von

Schüler*innen mit dem FALKE sowohl in fixierter, als auch computeradaptiver Version getestet werden können.

2.3.2 Testerleben

Unterschiede zwischen den genannten Testformaten von PPTs, CBTs und CATs können sich auf das Testerleben von Testpersonen auswirken (Chua, 2012; Colwell, 2013). Diese Dissertation befasst sich spezifisch mit kognitiven und affektiv-motivationalen Effekten. Im Folgenden wird auf die kognitive Belastung, die Testängstlichkeit und die Lesemotivation eingegangen. Sowohl Lesemotivation als auch Testängstlichkeit lassen sich in Zustands- und Eigenschaftsmerkmale einteilen (Tremblay et al., 1995; Zohar, 1998). Die Eigenschaft gibt in allgemeines Merkmal an, das die Prädisposition dafür widerspiegelt, Motivation oder Ängstlichkeit zu verspüren. Das Zustandsmerkmal hingegen gibt an, wie sich eine Person in einer spezifischen Situation fühlt.

Kognitive Belastung. Um geschriebenen Text zu verstehen, nutzen Leser*innen verschiedene kognitive Prozesse des Arbeits- und Langzeitgedächtnisses. Informationen werden im Arbeitsgedächtnis verarbeitet und im Langzeitgedächtnis gespeichert (Sweller et al., 2011). Beim *bottom-up processing* werden linguistische Signale wie Buchstaben oder Wörter erkannt, um eine Bedeutung zu finden. Beim *top-down processing* werden Erwartungen und Vorerfahrungen genutzt, um die Bedeutung von Wörtern zu erschließen (Goodman, 1970). Leser*innen nutzen beide Methoden gemeinsam, um Texte zu verstehen (Kintsch, 2005). Dabei wird angenommen, dass Repräsentationen des Gelesenen mithilfe des Arbeitsgedächtnisses erstellt werden (DeStefano & LeFevre, 2007). Im Arbeitsgedächtnis werden Vorerfahrungen top-down aus dem Langzeitgedächtnis abgerufen und bottom-up mit neuen Informationen aus dem Text zu einer Bedeutung integriert (Yang & Hu, 2022). Leseverständnis ist abhängig von der Kapazität des Arbeitsgedächtnisses (Schurer et al., 2020). Wenn die kognitiven Anforderungen eines Textes die Arbeitsgedächtniskapazität überschreiten, leidet das Textverständnis (Hahnel et al., 2019; Sweller, 2011). Die Kapazität des Arbeitsgedächtnisses kann durch kognitive Belastung beeinträchtigt werden (Sweller et al., 2011). Im Rahmen der c*ognitive load theory* (CLT; Sweller et al., 2011) wird zwischen verschiedenen Arten der kognitiven Belastung unterschieden. Intrinsische kognitive Belastung (ICL) bezieht sich auf die Belastung, die durch die Aufgabe direkt erzeugt wird. Dabei ist vor allem relevant, wie Teile eines Textes miteinander in Beziehung stehen (sog. *element interactivity*; Leppink, 2020). Wenn Elemente stark miteinander verknüpft sind, erhöht dies den Komplexitätsgrad des Textes, was wiederum zu erhöhter ICL führt (Sweller,

2011). Konkret auf Lesekompetenztests bezogen sollte die *element interactivity* eines Items mit dessen Schwierigkeit einhergehen (Noroozi & Karami, 2022), da schwierige Items eine höhere ICL erzeugen. Extrinsische kognitive Belastung (ECL) hängt mit der Darstellung einer Aufgabe zusammen. ECL ist eine zusätzliche Belastung, die durch die Darstellung der Aufgabe beziehungsweise des Textes entsteht. Die ECL beim Lesen eines Textes lässt sich direkt durch Entscheidungen im Design beeinflussen. So wird die ECL erhöht, wenn Texte verwaschen dargestellt werden (Eitel et al., 2014), da weitere kognitive Ressourcen eingesetzt werden müssen, um die Buchstaben, die undeutlicher dargestellt werden, zu identifizieren. Ein neuerer Aspekt der CLT ist der Ansatz der *working memory resource depletion* (O. Chen et al., 2018). Diese geht davon aus, dass die Kapazität des Arbeitsgedächtnisses eine limitierte Ressource ist, die bei Verwendung abnimmt. Dadurch kann eine Aufgabe, die das Arbeitsgedächtnis beansprucht, die Leistungsfähigkeit in zukünftigen ähnlichen oder identischen Aufgaben beeinträchtigen. Die Kapazität kann sich durch Ruhephasen wieder erholen (Tyler & Burns, 2008). Für Testsituationen legt dies nahe, dass anspruchsvolle Aufgaben die Leistung in späteren Aufgaben beeinflussen können.

Die Unterschiede zwischen papierbasierten, computerbasierten und computeradaptiven Tests können somit Auswirkungen auf die Arten der kognitiven Belastung im Rahmen der CLT haben. Weil sich Text- und Aufgabendarstellungen zwischen Papier und Computer unterscheiden (Gould, 1968), kann dies Auswirkungen auf die extrinsische kognitive Belastung haben.

Weiterhin können sich adaptive Tests potenziell auf die intrinsische Belastung auswirken. Nach Goldhammer et al. (2014) ist die Bearbeitungszeit, welche oftmals als Proxy-Variable für die kognitive Belastung genutzt wird (Leppink & Pérez-Fuster, 2019), eines Items im adaptiven Test abhängig sowohl von der Itemschwierigkeit als auch von der Personenfähigkeit. Die *element interactivity* eines Items in einem Test hängt mit dessen Itemschwierigkeit I_b zusammen (Noroozi & Karami, 2022).

Unter der Berücksichtigung, dass Personen mit einer hohen Fähigkeit auf vorherige Erfahrungen bei der Beantwortung von Aufgaben aus dem Langzeitgedächtnis zurückgreifen können (Yang & Hu, 2022), ist in diesem Kontext vor allem die Differenz der Fähigkeit einer Testperson (J) θ_j und der Itemschwierigkeit eines Items (I) I_b relevant. Aufgaben, deren Itemschwierigkeit dem Fähigkeitsniveau entsprechen ($\theta_j - I_b = 0$) würden so eine höhere intrinsische Belastung erzeugen, als Items, bei denen die Itemschwierigkeit deutlich unter dem Fähigkeitsniveau der Person liegt ($\theta_j - I_b < 0$), da die Testpersonen die Aufgaben mit

Vorerfahrungen oder im Bereich des Lesens durch automatisierte Prozesse bearbeiten können (Kalyuga, 2011; Yang & Hu, 2022). In computeradaptiven Tests werden diese Items, deren Schwierigkeit deutlich unter dem Fähigkeitsniveau der getesteten Person liegen, jedoch selten eingesetzt, da diese Items einen geringeren Informationsgehalt haben als anspruchsvollere Items. Im Gegensatz dazu sollten Items, deren Schwierigkeit deutlich über dem Fähigkeitsniveau der Testperson liegen ($\theta_j - I_b > 0$) mehr intrinsische Belastung erzeugen. Daher ist es möglich, dass die intrinsische kognitive Belastung in einem fixierten Test von der Fähigkeit der Testperson abhängt, im adaptiven Test jedoch über die Fähigkeitsniveaus gleichbleibt.

Testängstlichkeit. Testängstlichkeit ist ein wichtiges motivationales Konstrukt. Sie ist definiert als die physiologische, behaviorale und kognitive Reaktion auf einen Test oder eine ähnliche evaluative Situation (Zeidner, 1998). Eine physiologische Reaktion ist beispielweise physische Anspannung, kognitive Reaktionen umspannen Besorgtheit und ablenkende Gedanken und behavoriale Reaktionen beschreiben vermeidendes Verhalten gegenüber Testsituationen. Diese Dissertation nutzt die Zwei-Komponenten Theorie (Liebert & Morris, 1967), nach der sich Testängstlichkeit in eine Besorgtheits- (worry) und eine Aufgeregtheitskomponente (emotionality) aufteilen lässt. Besorgtheit bezeichnet kognitive Aspekte, wie Sorgen über mögliche Folgen nach einem schlechten Testergebnis oder Vergleiche mit Mitschüler*innen, und Aufgeregtheit physiologische Prozesse, wie Angespanntheit.

Im additiven Modell der Testängstlichkeit, besteht diese aus einer Zustands- und eine Eigenschaftskomponente (Zohar, 1998). Testängstlichkeit als Eigenschaft beschreibt, inwiefern eine Person allgemein und außerhalb von Testsituationen dazu neigt, Testängstlichkeit zu verspüren. Zustandsangst beschreibt hingegen die Testangst, die akut in einer Testsituation verspürt wird. Im additiven Modell wird die Zustandsangst von der allgemeinen Testängstlichkeit und Zustandsvariablen beeinflusst. Eigenschafts- und Zustandsangst sind daher hoch korreliert (Bertrams et al., 2010).

Testängstlichkeit und Testleistungen korrelieren negativ miteinander (Cassady & Johnson, 2002; Zeidner, 1998). Dabei gibt es verschiedene Annahmen darüber, in welchem kausalen Zusammenhang die Testleistung und die Testängstlichkeit stehen. Im Interferenzmodell wird davon ausgegangen, dass Testpersonen, die eine hohe Testängstlichkeit verspüren, aufgrund dieser Angst von der Bearbeitung der Testaufgaben abgelenkt werden, wodurch sich die Leistung verringert (Eysenck et al., 2007). Dem

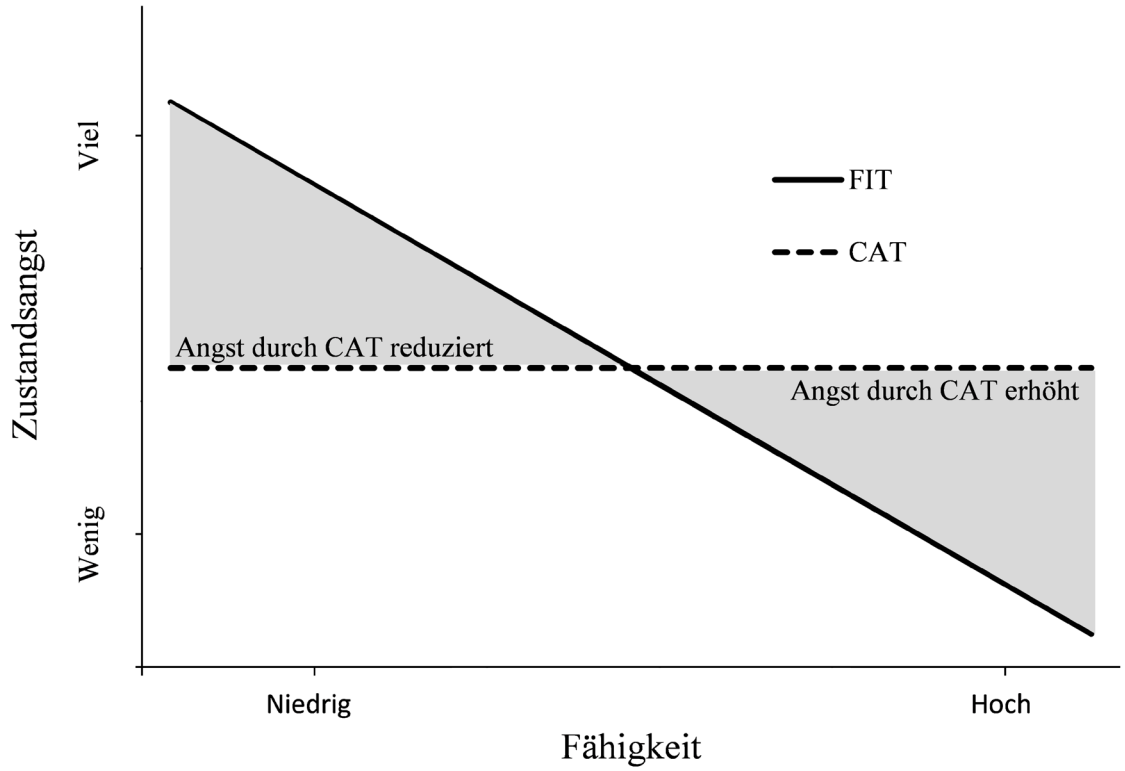
gegenüber steht das Defizitmodell, in dem angenommen wird, dass Personen mit einer niedrigen Fähigkeit mehr Testängstlichkeit verspüren, da sie sich ihrer mangelnden Fähigkeit bewusst sind (Sommer & Arendasy, 2014). Zuletzt gibt es die Annahme einer reziproken Beziehung zwischen den Variablen (Steinmayr et al., 2015) in der Testängstlichkeit die Leistung beeinträchtigt, wodurch wiederum mehr Testängstlichkeit generiert wird (Zeidner, 1998).

Testängstlichkeit beim Testmedium. Der Einsatz von computerbasierten Tests in Leistungssituationen setzt voraus, dass Testpersonen in der Lage sind, mit dem digitalen Endgerät umzugehen. Die Erwartung und Selbsteinschätzung von Testpersonen an ihre eigene Kompetenz, den Computer für die Testsituation adäquat zu bedienen, wird als Computerselbstwirksamkeit bezeichnet (Moos & Azevedo, 2009). Vor allem für ältere Testpersonen ist die Computerselbstwirksamkeit ein wichtiger Faktor für die verspürte Testängstlichkeit bei einem computerbasierten Test (I.-S. Chen, 2017). Personen, die sich selbst als kompetent mit dem Umgang mit Computern ansehen, machen sich weniger Sorgen darüber, ob sie mit dem Computer für die Aufgabenstellung hinreichend umgehen können. Daher ist die Computerangst in diesem Kontext grundsätzlich ein relevantes Konstrukt (Cassady & Gridley, 2005). Testpersonen haben eine Selbstwirksamkeitserwartung, wie gut sie die Anforderung an ihre computerbezogenen Fähigkeiten umsetzen können (Saadé & Kira, 2009). Wenn Testpersonen aufgrund einer mangelnden Selbstwirksamkeit mit digitalen Medien vermuten, dass sie den Anforderungen an ihre computerbezogenen Fähigkeiten nicht gerecht werden können und somit Computerangst verspüren, kann sich dies auf ihre Zustandsangst auswirken. Shermis und Lombard (1998) argumentieren weiterhin, dass Computerangst eine Manifestation von Testängstlichkeit ist, wenn ein Test am Computer durchgeführt wird. Allerdings ist zu beachten, dass ältere Personen im Durchschnitt mehr Computerangst verspüren als jüngere Personen (Powell, 2013). Entsprechend fanden Fritts und Marszalek (2010) in einer Studie mit Schüler*innen der sechsten bis achten Klasse (M_{Alter} = 13.36 Jahre, SD_{Alter} = 1.15 Jahre) auf einem Testängstlichkeitsindex mit einer Reichweite von 10 bis 50 Mittelwerte von 17.86 (SD = 0.66) und auch Powell (2013) vermutete, dass Computerangst für Personen, die umgeben von digitalen Medien aufwachsen, zurückgehen würde. Vor diesem Hintergrund ist es möglich, dass sich der Testmodus Papier beziehungsweise Computer über die Computerangst auf die Testängstlichkeit von Schüler*innen auswirken kann (Shermis & Lombard, 1998). Dabei können die Effekte in der Grundschule im Vergleich zu älteren Testpersonen geringer ausfallen.

Testängstlichkeit beim Testmodus. Ein wichtiger Bestandteil von computeradaptiven Tests ist die fixierte Lösungsrate, die bei 1PL- und 2PL-Modellen bei etwa 50 Prozent liegt. Diese Lösungsrate kann dafür sorgen, dass Testpersonen einen CAT als deutlich schwieriger wahrnehmen als einen äquivalenten FIT (Ling et al., 2017). Vor allem Schüler*innen in der Primar- und Sekundarstufe können davon betroffen sein, da Tests in der Schule oft so konstruiert sind, dass ein*e durchschnittliche*r Schüler*in eine höhere durchschnittliche Lösungsrate hat (Eggen & Verschoor, 2006). In einem CAT kann so für Schüler*innen eine Diskrepanz zwischen der Lösungsrate, die Schüler*innen von sich selbst erwarten, und der Lösungsrate, die sie selbst im Test wahrnehmen, entstehen. Diese Diskrepanz kann sich auf die Erfolgserwartung der Schüler*innen im Test und somit auch auf die Zustandsangst auswirken (Ortner & Caspers, 2011). Dies gilt sowohl für leistungsstarke Schüler*innen, deren Erfolgserwartungen nicht erfüllt werden, als auch für leistungsschwache Schüler*innen, die ihre Erfolgserwartungen in einem CAT übertreffen können. Verschiedene Studien untersuchten bereits diesen Zusammenhang, mit heterogenen Ergebnissen. Ortner et al. (2014) zeigten eine erhöhte Zustandsangst in einem Test mit Oberstufenschüler*innen aus Deutschland zum räumlichen Denken. Ling et al. (2017) verglichen zusätzlich zu einem FIT und einem CAT mit Lösungsraten von 50 Prozent einen weiteren CAT mit einer Lösungsrate von 70 Prozent und kamen zu dem Ergebnis, dass die verspürte Testängstlichkeit von Schüler*innen der sechsten bis achten Klasse an einer amerikanischen Middle School im Durchschnitt am höchsten im CAT mit einer Lösungsrate von 50 Prozent war. Dem gegenüber steht eine Studie mit einem *multistage-adaptive test* (MAT) von Martin und Lazendic (2018) an mehreren australischen Schulen, bei denen Schüler*innen der dritten, fünften, siebten und neunten Klasse keine Unterschiede in ihrer durchschnittlichen angegebenen Testängstlichkeit in einem Mathematiktest beschrieben. Besonders Schüler*innen, die leistungsstark oder leistungsschwach sind, könnten von der Diskrepanz zwischen erwarteter und erfahrener Lösungsrate in einem CAT betroffen sein. Starke Schüler*innen würden weniger und schwache Schüler*innen mehr Aufgaben korrekt beantworten können, als sie es von anderen Tests gewöhnt sind. Im Rahmen des Interferenzmodells der Testängstlichkeit (Eysenck et al., 2007) können diese Schüler*innen dann mehr beziehungsweise weniger Testängstlichkeit in einem CAT verspüren als in einem FIT. Abbildung 1 stellt diese theoretische Beziehung zwischen der Testängstlichkeit und der Fähigkeit in einem FIT und einem CAT dar.

Abbildung 1

Theoretisches Model der Interaktion von Zustandsangst und Testleistung in einem FIT und einem CAT.



Zu beachten ist, dass die bisherige Befundlage zum Effekt von CAT auf die Testängstlichkeit sich vor allem auf den mathematischen und räumlichen Bereich mit älteren Schüler*innen fokussiert. Vor allem in der Primarstufe ist der Effekt noch nicht hinreichend untersucht. Gerade für jüngere Schüler*innen, die aufgrund ihres Alters noch weniger Erfahrung mit Testsituationen hatten, wäre es wichtig zu untersuchen, ob sich auch hier eine erhöhte Testängstlichkeit findet.

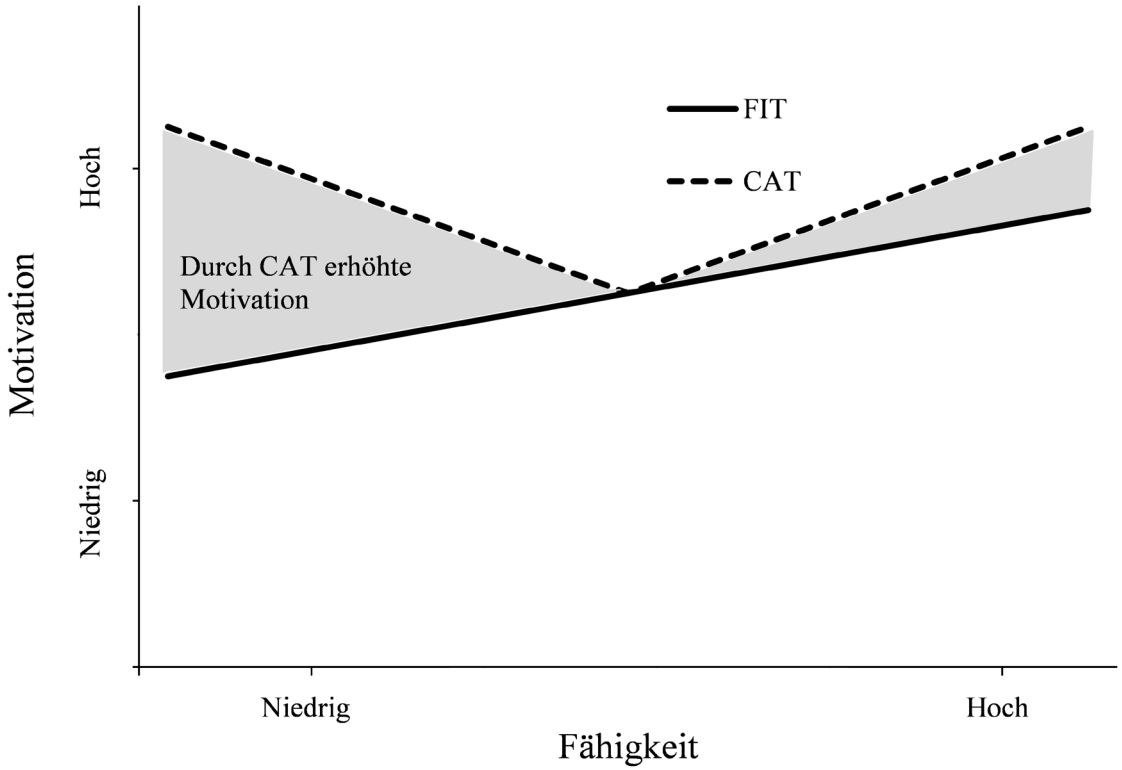
Lesemotivation. Lesemotivation ist ein Prädiktor für Testleistung in einem Lesekompetenztest. Dieser Zusammenhang ist theoretisch im Rahmen der Erwartungs-Wert Theorie (Wigfield, 1994) begründet, da die Motivation einen Aspekt der Werte abbildet, die sich auf den akademischen Erfolg auswirken. Empirisch gibt es ebenfalls klare Befunde, dass die Testleistung von der Lesemotivation abhängt (Becker et al., 2010; McElvany & Schwabe, 2019). Dabei lässt sich die Lesemotivation, wie auch zuvor die Testängstlichkeit, in Zustands- und Eigenschaftskomponenten aufteilen (Tremblay et al., 1995). Die allgemeine Lesemotivation gibt an, inwiefern eine Person grundsätzlich motiviert ist, Texte zu lesen, während die situative Lesemotivation autonym situationsspezifisch wirkt.

Lesemotivation und digitale Medien. Ein theoretischer Ansatz im Bereich der Anwendung digitaler Medien im Bildungsbereich ist der *novelty effect* (Clark, 1983; Shin et al., 2019). Dieser besagt, dass neue Stimuli *aufgrund ihrer Neuheit* die intrinsische Motivation (Lomas et al., 2017) und sogar den Lernerfolg erhöhen können (Burke & James, 2008). Dieser motivierende Effekt sinkt jedoch mit wachsender Gewöhnung an das Medium (Keller & Suzuki, 2004). Digitale Medien im Bildungsbereich leiden oft unter diesem Effekt, da Nutzer*innen nur kurzfristig motiviert sind, wodurch die Nutzungshäufigkeit ebenfalls abnimmt (Rodrigues et al., 2022; Shin et al., 2019). Mehrere Studien, die das Lesemedium von Leser*innen zwischen Papier und Computern verglichen, kamen zu dem Ergebnis, dass digitale Medien zu einer erhöhten Motivation führen, sowohl beim Lesen im Privaten (Picton, 2014), als auch in Testsituationen (Chua, 2012). Mehrere Studien konnten ebenfalls zeigen, dass Schüler*innen digitale Medien zum Lesen dem Papier gegenüber bevorzugen (Golan et al., 2018; Tveit & Mangen, 2014). Aufgrund des *novelty effects* ist es naheliegend, dass die motivierenden Effekte von digitalen Medien beim Lesen auch im Schulkontext nicht langfristig anhalten.

Bei computeradaptiven Tests nimmt die Motivation eine spezielle Rolle ein. Schon Weiss und Betz (1973) schlugen vor, dass CATs auf Testpersonen allgemein motivierend wirken würden, da Personen mit niedriger Fähigkeit durch den einfacheren Test weniger entmutigt werden, während Personen mit hoher Fähigkeit durch die anspruchsvolleren Aufgaben gefördert werden und sich weniger langweilen. Die Annahme ist in Abbildung 2 dargestellt.

Abbildung 2

Schematische Darstellung der Motivation in einem CAT gegenüber einem FIT nach Weiss und Betz, 1973



In der Praxis ließ sich diese theoretische Annahme jedoch bisher nicht eindeutig bestätigen. Weder Ling et al. (2017) noch Martin & Lazendic (2018) konnten Unterschiede in der Motivation von Schüler*innen nach einem computerbasierten oder computeradaptiven Mathematiktest feststellen. Es gilt noch zu untersuchen, ob die Annahme in einem Lesekompetenztest zu finden ist.

2.4 Zentrale Forschungsanliegen

Im Folgenden werden die zentralen Forschungsanliegen, die sich aus den oben genannten theoretischen Rahmenbedingungen ergeben, erläutert. Die Rückstände, die sich in Deutschland im Bereich der Digitalisierung an Schulen vor den COVID-19-Pandemie beschreiben ließen (Fraillon et al., 2019), wirkten sich stark auf den Unterricht während den Schulschließungen aus (Freundl et al., 2021). Besonders unter der Berücksichtigung des TAM (Davis, 1985) und den Einflüssen von Einstellungen und Selbstwirksamkeitserwartungen auf die Nutzung von digitalen Medien im Unterricht (Kihoza et al., 2016; Paraskeva et al., 2008) ist von Interesse, wie sich die Einstellungen von Lehrkräften gegenüber digitalen Medien und

ihre Selbstwirksamkeit im Rahmen des TPACK-Modells (Koehler & Mishra, 2009) zwischen 2020 und 2021 verändert haben. Diese erste Forschungsfrage wird in Beitrag I bearbeitet. Berücksichtigt werden vor allem, wie sich die technischen Grundvoraussetzungen wandelten, wie zwischen Lehrkräften und Schüler*innen kommuniziert wurde und inwiefern digitale Medien für den Unterricht im Verlauf der Pandemie eingesetzt wurden. Weiterhin wurde untersucht, inwiefern sich die Selbsteinschätzungen der Lehrkräfte im Rahmen des TPACK-Modells gewandelt haben und wie Lehrkräfte die Effekte von digitalen Medien auf ihre Schüler*innen einschätzten.

Auf der Ebene der Schüler*innen wurden die Auswirkungen von digitalen Medien auf kognitive und affektiv-motivationale Zustände während Testsituationen betrachtet, die im Unterricht zur Kompetenzdiagnose genutzt werden können. Dabei wurde untersucht, wie sich die Substitution von analogen Papiertests und die Augmentation von computerbasierten Testverfahren mithilfe von computeradaptiven Tests auf das Testerleben von Schüler*innen auswirkt. Obgleich es eine substanzielle Forschungsbasis für Unterschiede zwischen PPTs, CBTs und CATs gibt, wurden in bisherigen Studien nur jeweils zwei der drei Formate miteinander verglichen. Vor allem bei Vergleichen zwischen PPTs und CATs können so konfundierende Effekte des Mediums auf Vergleiche des Modus wirken (Ling et al., 2017). Im deutschen Kontext, in dem digitale Medien kaum zur Kompetenzerfassung eingesetzt wurden (Fraillon et al., 2019), ist es daher relevant, Unterschiede zwischen PPTs, CBTs und CATs simultan zu untersuchen. Weiterhin gibt es noch wenig Forschung zu Effekten des Modus und des Mediums im Grundschulbereich. Gerade jüngere Schüler*innen haben weniger Erfahrung mit Testsituationen und wachsen umgeben von digitalen Medien auf (Fraillon et al., 2019).

Insbesondere die kognitive Belastung ist ein wichtiger Anhaltspunkt, um Unterschiede zwischen den Formaten zu erklären. Darstellungsunterschiede zwischen Papier und Bildschirm können sich auf die extrinsische Belastung auswirken und Unterschiede in der Aufgabenabfolge bei CBTs und CATs können sich auf die intrinsische Belastung auswirken. Unter Berücksichtigung der CLT (Sweller, 2011) wird in Beitrag II der Fragestellung nachgegangen, welche Effekte papierbasierte, computerbasierte und computeradaptive Tests auf die kognitive Belastung von Schüler*innen der vierten Klasse in einem Lesekompetenztest haben. Vertiefend dazu werden die motivational-affektiven Merkmale der Lesemotivation und Testängstlichkeit in Beitrag III untersucht. Während die Testängstlichkeit bei adaptiven Tests ein zentrales Forschungsthema in der Literatur darstellt (Ling et al., 2017;

Ortner & Caspers, 2011), mangelt es an Befunden über jüngere Schüler*innen, vor allem im Bereich des Lesens. Daher wird neben der kognitiven Belastung in der Grundschule in Beitrag II vertiefend auf die spezifischen Effekte der Testformate auf die Lesemotivation und Testängstlichkeit in Beitrag III eingegangen.

Das übergeordnete Forschungsanliegen liegt somit darin tiefergehend zu ergründen, wie sich der Einsatz digitaler Medien im Unterricht auf die Einstellungen und Erfahrungen von sowohl Lehrkräften als auch Schüler*innen auswirken.

Literaturverzeichnis I

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education*, *126*, 334–345. https://doi.org/10.1016/j.compedu.2018.07.021

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and Challenges of Using e-Assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37. https://doi.org/10.18178/ijiet.2018.8.1.1008

Baier, D., & Kamenowski, M. (2020). *Wie erlebten Jugendliche den Corona-Lockdown?: Ergebnisse einer Befragung im Kanton Zürich*. Zürich: Züricher Fachhochschule für Angewandte Wissenschaften.

Becker, M., McElvany, N., & Kortenbruck, M. (2010). Intrinsic and extrinsic reading motivation as predictors of reading literacy: A longitudinal study. *Journal of Educational Psychology*, *102*(4), 773–785. https://doi.org/10.1037/a0020084

Bertrams, A., Englert, C., & Dickhäuser, O. (2010). Self-control strength in the relation between trait test anxiety and state anxiety. *Journal of Research in Personality*, *44*(6), 738–741. https://doi.org/10.1016/j.jrp.2010.09.005

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

BMG. (2022). *Coronavirus-Pandemie: Was geschah wann?* Bundesministerium für Gesundheit. https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html

Burke, L. A., & James, K. E. (2008). Powerpoint-Based Lectures in Business Education: An Empirical Investigation of Student-Perceived Novelty and Effectiveness. *Business Communication Quarterly*, *71*(3), 277–296. https://doi.org/10.1177/1080569908317151

Cassady, J. C., & Gridley, B. E. (2005). The Effects of Online Formative and Summative Assessment on Test Anxiety and Performance. *The Journal of Technology, Learning and Assessment*, *4*(1).

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295.

Cress, U., Diethelm, I., Eickelmann, B., Köller, O., Nickolaus, R., Pant, H. A., & Reiss, K. (2018). Schule in der digitalen Transformation. *Perspektiven der Bildungswissenschaften.*

Chen, I.-S. (2017). Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behavior*, *72*, 362–370. https://doi.org/10.1016/j.chb.2017.02.059

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educational Psychology Review*, *30*(2), 483–501. https://doi.org/10.1007/s10648-017-9426-2

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, *28*(5), 1580–1586. https://doi.org/10.1016/j.chb.2012.03.020

Clark, R. E. (1983). Reconsidering Research on Learning from Media. *Review of Educational Research*, *53*(4), 445–459. https://doi.org/10.3102/00346543053004445

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies*, *1*(2), 50–60.

Cruz-Cárdenas, J., Zabelina, E., Guadalupe-Lanas, J., Palacio-Fierro, A., & Ramos-Galarza, C. (2021). Covid-19, consumer behavior, technology, and society: A literature review and bibliometric analysis. *Technological Forecasting and Social Change*, *173*, 121179. https://doi.org/10.1016/j.techfore.2021.121179

Culpepper, S. A. (2016). Revisiting the 4-Parameter Item Response Model: Bayesian Estimation and Application. *Psychometrika*, *81*(4), 1142–1163. https://doi.org/10.1007/s11336-015-9477-6

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems. *Council of Chief State School Officers*.

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*.

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. https://doi.org/10.1016/j.edurev.2018.09.003

DeStefano, D., & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, *23*(3), 1616–1641.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. *Applied Psychological Measurement*, *30*(5), 379–393. https://doi.org/10.1177/0146621606288890

Ehrich, S. (2002). On stratified extensions of Gauss-Laguere and Gauss-Hermite quadrature formulas. *Journal of Computational and Applied Mathematics*, *140*(1-2), 291–299.

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., & Vahrenhold, J. (2019). *ICILS 2018 #Deutschland: Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking*. Waxmann.

Eickelmann, B., & Gerick, J. (2020). Lernen mit digitalen Medien. Zielsetzungen in Zeiten von Corona und unter besonderer Berücksichtigung von sozialen Ungleichheiten. In D. Fickermann & B. Edelstein (Hrsg.), *„Langsam vermisse ich die Schule…". Schule während und nach der Corona-Pandemie.* (pp. 153–162). Waxmann.

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency Meets Cognitive Load in Multimedia Learning: Does Harder-to-Read Mean Better-to-Understand? *Applied Cognitive Psychology*, *28*(4), 488–501. https://doi.org/10.1002/acp.3004

Embretson, S. E., & Reise, S. P. (2009). *Item Response Theory for Pychologists*. Psychology Press.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fend, H. (2009). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen*. Springer-Verlag.

Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & Beurs, E. de (2016). Simulating computer adaptive testing with the Mood and Anxiety Symptom Questionnaire. *Psychological Assessment*, *28*(8), 953–962. https://doi.org/10.1037/pas0000240

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2019). *Preparing for life in a digital world: Iea international computer and information literacy study 2018 international report: Iea international computer and information*. Springer.

Freundl, V., Stiegler, C., & Zierow, L. (2021). Europas Schulen in der Corona-Pandemie – ein Ländervergleich. *ifo Schnelldienst*, *74*(12), 41–50.

Fritts, B. E., & Marszalek, J. M. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education*, *13*(3), 441–458.

Furenes, M. I., Kucirkova, N., & Bus, A. G. (2021). A Comparison of Children's Reading on Paper Versus Screen: A Meta-Analysis. Review of Educational Research, 91(4), 483–517. https://doi.org/10.3102/0034654321998074

Golan, D. D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. Computers & Education, 126, 346–358. https://doi.org/10.1016/j.compedu.2018.08.001

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H. & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. Journal of Educational Psychology, 106(3), 608–626. https://doi.org/10.1037/a0034716

Goodman, K. S. (1970). Psycholinguistic universals in the reading process. *Visible Language*, *4*(2), 103-110.

Gould, J. D. (1968). Visual factors in the design of computer-controlled CRT displays. *Human Factors*, *10*(4), 359–375. https://doi.org/10.1177/001872086801000406

Hahnel, C., Schoor, C., Kroehne, U., Goldhammer, F., Mahlow, N., & Artelt, C. (2019). The role of cognitive load in university students' comprehension of multiple documents. *Zeitschrift Für Pädagogische Psychologie*, *33*(2), 105–118. https://doi.org/10.1024/1010-0652/a000238

Hamilton, E. R., Rosenberg, J. M., & Akcaoglu, M. (2016). The Substitution Augmentation Modification Redefinition (SAMR) Model: A Critical Review and Suggestions for its Use. *TechTrends*, *60*(5), 433–441. https://doi.org/10.1007/s11528-016-0091-y

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hauck-Thum (2018). Fachspezifische Möglichkeiten und Potentiale von Adaptable Books im Lese- und Literaturunterricht der Grundschule. *Mitteilung Des Deutschen Germanistenverbandes*, *65*, 294–305.

Helm, C., Huber, S., & Loisinger, T. (2021). Was wissen wir über schulische Lehr-Lern-Prozesse im Distanzunterricht während der Corona-Pandemie? – Evidenz aus Deutschland, Österreich und der Schweiz . *Zeitschrift für Erziehungswissenschaft*, *24*(2), 237–311. https://doi.org/10.1007/s11618-021-01000-z

Helm, C., & Warwas, J. (2018). Psychological determinants of test motivation in low-stakes test situations: A longitudinal study of singletrait–multistate models in accounting. *Empirical Research in Vocational Education and Training*, *10*(1), 1–34. https://doi.org/10.1186/s40461-018-0074-7

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster; New York: Waxmann.

Jude, N., Ziehm, J., Goldhammer, F., Drachsler, H., & Hasselhorn, M. (2020). *Digitalisierung an Schulen–eine Bestandsaufnahme*. Frankfurt am Main: DIPF| Leibniz-Institut für Bildungsforschung und Bildungsinformation.

Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, *23*(1), 1–19. https://doi.org/10.1007/s10648-010-9150-7

Keller, J., & Suzuki, K. (2004). Learner motivation and E-learning design: A multinationally validated process. *Journal of Educational Media*, *29*(3), 229–239. https://doi.org/10.1080/1358165042000283084

Kihoza, P., Zlotnikova, I., Bada, J., & Kalegele, K. (2016). Classroom ICT integration in Tanzania: Opportunities and challenges from the perspectives of TPACK and SAMR models. *International Journal of Education and Development Using ICT*, *12*(1).

Kintsch, W. (2005). An Overview of Top-Down and Bottom-Up Effects in Comprehension: The CI Perspective. *Discourse Processes*, *39*(2-3), 125–128. https://doi.org/10.1080/0163853X.2005.9651676

KMK [Kultusministerkonferenz]. (2016). *Bildung in der digitalen Welt: Strategie der Kultusministerkonferenz*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

KMK [Kultusministerkonferenz]. (2021). *Lehren und Lernen in der digitalen Welt: Ergänzung zur Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Koehler, M., & Mishra, P. (2009). What is Technological Pedagogical Content Knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education*, *9*(1), 60–70.

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, *123*, 138–149.

Köpper, M., Mayr, S., & Buchner, A. (2016). Reading from computer screen versus reading from paper: Does it still make a difference? *Ergonomics*, *59*(5), 615–632. https://doi.org/10.1080/00140139.2015.1100757

Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of Screen Versus Print Reading Comprehension Depends on Task Complexity and Proficiency. *Discourse Processes*, *54*(5-6), 427–445. https://doi.org/10.1080/0163853X.2017.1319653

Leppink, J. (2020). Revisiting cognitive load theory: Second thoughts and unaddressed questions. *Scientia Medica*, *30*(1), e36918. https://doi.org/10.15448/1980-6108.2020.1.36918

Leppink, J. & Pérez-Fuster, P. (2019). Mental Effort, Workload, Time on Task, and Certainty: Beyond Linear Models. *Educational Psychology Review*, *31*(2), 421–438. https://doi.org/10.1007/s10648-018-09460-2

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, *20*(3), 975–978. https://doi.org/10.2466/pr0.1967.20.3.975

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511.

Lomas, J. D., Koedinger, K., Patel, N., Shodhan, S., Poonwala, N., & Forlizzi, J. L. (2017). Is Difficulty Overrated? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3025453.3025638

Lorenz, R., Brüggemann, T., Eickelmann, B., & McElvany, N. (2022). Gelingensbedingungenfür dem Einsatz digtialer Medien in Lernsituationenin der Grundschule im Bereich Lesen - Befunde einer qualitativen Befragung von Lehrpersonen. In F. Lauermann, C. Jöhren, N. McElvany, M. Becker, & H. Gaspard (Hrsg.), *Jahrbuch der Schulentwicklung Band 22: Multiperspektivität von Unterrichtsprozessen* (pp. 65–93). Beltz Juventa.

Lorenz, R., Yotyodying, S., Eickelmann, B., & Endberg, M. (2022). *Schule digital - der Länderindikator 2021: Lehren und Lernen mit digitalen Medien in der Sekundarstufe I in Deutschland im Bundesländervergleich und im Trend seit 2017* (1. Auflage). Waxmann.

Ludewig, U., Kleinkorres, R., Schaufelberger, R., Schlitter, T., Lorenz, R., König, C., ... & McElvany, N. (2022). Covid-19 pandemic and student reading achievement: Findings from a school panel study. *Frontiers in Psychology*, *13*.

Ludewig, U., Trendtel, M., Schlitter, T., & McElvany, N. (2021). Adaptives Testen von Textverständnis in der Grundschule. *Diagnostica, 68*(1), 39-50.

Lundh, F. (1999). An introduction to tkinter. www.Pythonware.Com/Library/Tkinter/Introduction/Index.Html

Mangen, A., & Kuiken, D. (2014). Lost in an iPad. *Scientific Study of Literature*, *4*(2), 150–177. https://doi.org/10.1075/ssol.4.2.02man

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society*, *14*(1), 81–95. https://doi.org/10.1007/s10209-014-0348-1

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27.

Mayr, S., Köpper, M., & Buchner, A. (2017). Effects of high pixel density on reading comprehension, proofreading performance, mood state, and physical discomfort. *Displays*, *48*, 41–49. https://doi.org/10.1016/j.displa.2017.03.002

McElvany, N., & Schwabe, F. (2019). Gender gap in reading digitally? Examining the role of motivation and self-concept. *Journal for Educational Research Online*, *11*(1), 145–165. https://doi.org/10.25656/01:16791

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the Python in Science Conference, Proceedings of the 9th Python in Science Conference* (pp. 56–61). SciPy. https://doi.org/10.25080/Majora-92bf1922-00a

Michel, P., Baumstarck, K., Loundou, A., Ghattas, B., Auquier, P., & Boyer, L. (2018). Computerized adaptive testing with decision regression trees: An alternative to item response theory for quality of life measurement in multiple sclerosis. *Patient Preference and Adherence*, *12*, 1043–1053. https://doi.org/10.2147/PPA.S162206

Mishra, P., & Koehler, M. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Moos, D. C., & Azevedo, R. (2009). Learning With Computer-Based Learning Environments: A Literature Review of Computer Self-Efficacy. *Review of Educational Research*, *79*(2), 576–600. https://doi.org/10.3102/0034654308326083

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In *Testtheorie und Fragebogenkonstruktion* (pp. 227–274). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-20072-4_10

MSB NRW [Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen]. (2022). *Schulbetrieb im Wechselunterricht ab Montag, 19. April 2021.* Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen.

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia*, *12*(1), 1–19. https://doi.org/10.1186/s40468-022-00163-8

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375.

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*.

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I Will Probably Fail. *European Journal of Psychological Assessment*. https://econtent.hogrefe.com/doi/full/10.1027/1015-5759/a000168

Paraskeva, F., Bouta, H., & Papagianni, A. (2008). Individual characteristics and computer self-efficacy in secondary education teachers to integrate technology in educational practice. *Computers & Education*, *50*(3), 1084–1091. https://doi.org/10.1016/j.compedu.2006.10.006

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education*, *2*(1), 1–17. https://doi.org/10.1186/s40536-014-0005-4

Petko, D. (2012). Teachers' pedagogical beliefs and their use of digital media in classrooms: Sharpening the focus of the 'will, skill, tool' model and integrating teachers' constructivist orientations. *Computers & Education*, *58*(4), 1351–1359. https://doi.org/10.1016/j.compedu.2011.12.013

Picton, I. (2014). The Impact of eBooks on the Reading Motivation and Reading Skills of Children and Young People: A Rapid Literature Review. *National Literacy Trust*.

Piolat, A., Roussey, J.-Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, *47*(4), 565–589. https://doi.org/10.1006/ijhc.1997.0145

Powell, A. L. (2013). Computer anxiety: Comparison of research from the 1990s and 2000s. *Computers in Human Behavior*, *29*(6), 2337–2381. https://doi.org/10.1016/j.chb.2013.05.012

Puentedura, R. (2006). Transformation, technology, and education [Blog post]. Abgerufen am 22.02.2023 von http://hippasus.com/resources/tte/.

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, *2*(5), 128-137.

Rodrigues, L., Pereira, F. D., Toda, A. M., Palomino, P. T., Pessoa, M., Carvalho, L. S. G., Fernandes, D., Oliveira, E. H. T., Cristea, A. I., & Isotani, S. (2022). Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International Journal of Educational Technology in Higher Education*, *19*(1), 1–25. https://doi.org/10.1186/s41239-021-00314-6

Rowsell, J., & Burke, A. (2009). Reading by Design: Two Case Studies of Digital Reading Practices. *Journal of Adolescent & Adult Literacy*, *53*(2), 106–118. https://doi.org/10.1598/JAAL.53.2.2

Saadé, R., & Kira, D. (2009). Computer Anxiety in E-Learning: The Effect of Computer Self-Efficacy. *Journal of Information Technology Education: Research*, *8*(1), 177–191.

Schurer, T., Opitz, B., & Schubert, T. (2020). Working memory capacity but not prior knowledge impact on readers' attention and text comprehension. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, *14*(1), 111–123. https://doi.org/10.1016/S0747-5632(97)00035-6

Shin, G., Feng, Y., Jarrahi, M. H., & Gafinowitz, N. (2019). Beyond novelty effect: A mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open*, *2*(1), 62–72. https://doi.org/10.1093/jamiaopen/ooy048

Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence*, *42*, 115–127. https://doi.org/10.1016/j.intell.2013.11.003

Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2015). Subjective Well-Being, Test Anxiety, Academic Achievement: Testing for Reciprocal Effects. *Frontiers in Psychology*, *6*, 1994. https://doi.org/10.3389/fpsyg.2015.01994

Sweller, J. (2011). Cognitive Load Theory. In *Psychology of Learning and Motivation* (pp. 37–76). Elsevier. https://doi.org/10.1016/b978-0-12-387691-1.00002-8

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring Cognitive Load. In *Cognitive Load Theory* (pp. 71–85). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-8126-4_6

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

SWK [Ständige Wissenschaftliche Kommission]. (2022). *Digitalisierung im Bildungssystem: Handlungsempfehlungen von der Kita bis zur Hochschule*. Ständige Wissenschaftliche Kommission der Kultusministerkonferenz.

Tremblay, P. F., Goldberg, M. P., & Gardner, R. C. (1995). Trait and state motivation and the acquisition of Hebrew vocabulary. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *27*(3), 356–370. https://doi.org/10.1037/0008-400X.27.3.356

Trültzsch-Wijnen, C., & Trültzsch-Wijnen, S. (2020). Remote Schooling during the Covid-19 Lockdown In Austria (Spring 2020). *KiDiCoTi National Report.* https://doi.org/10.25598/KiDiCoTi-AT-2020-1.

Tveit, Å. K., & Mangen, A. (2014). A joker in the class: Teenage readers' attitudes and preferences to reading on different devices. *Library & Information Science Research*, *36*(3-4), 179–184. https://doi.org/10.1016/j.lisr.2014.08.001

Tyler, J. M., & Burns, K. C. (2008). After Depletion: The Replenishment of the Self's Regulatory Resources. *Self and Identity*, *7*(3), 305–321. https://doi.org/10.1080/15298860701799997

Ukonu, M. O., Ohaja, E. U., Okeke, S. V., & Okwumbu, R. O. (2021). Interactive effects of institutional requirements and screen vs. Print platforms on preference of Times New Roman and Calibri among university students. *Cogent Education*, *8*(1), Article 1968779. https://doi.org/10.1080/2331186X.2021.1968779

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238.

Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or Adaptive?* (Research Rep. No 73-1). Minneapolis: University of Minnesota, Psychometric Methods Program.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49–78. https://doi.org/10.1007/BF02209024

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Oecd Education Working Papers* (Vol. 209). OECD.

Yang, X., & Hu, J. (2022). Distinctions between mobile-assisted and paper-based EFL reading comprehension performance: Reading cognitive load as a mediator. *Computer Assisted Language Learning*, 1–32. https://doi.org/10.1080/09588221.2022.2143527

Zeidner, M. (1998). *Test anxiety: The state of the art. Perspectives on individual differences.* Kluwer Academic/Plenum Publishers.

Zheng, Y., Cheon, H., & Katz, C. M. (2020). Using Machine Learning Methods to Develop a Short Tree-Based Adaptive Classification Test: Case Study With a High-Dimensional Item Pool and Imbalanced Data. *Applied Psychological Measurement*, *44*(7-8), 499–514. https://doi.org/10.1177/0146621620931198

Zohar, D. (1998). An additive model of test anxiety: Role of exam-specific expectations. *Journal of Educational Psychology*, *90*(2), 330–340. https://doi.org/10.1037/0022-0663.90.2.330

3. Beiträge der kumulativen Promotion

3.1 Beitrag I: Unterricht zu Beginn und nach einem Jahr der Corona-Pandemie – Lehrkräftebefragungen zum Lernen mit digitalen Medien im Vergleich
Welche Unterschiede werden für den Unterricht während den Schulschließungen 2020 und 2021 deutlich?

Lorenz, R., **Brüggemann, T**., Stang-Rabrig, J., McElvany, N. (2023). Unterricht zu Beginn und nach einem Jahr der Corona-Pandemie – Lehrkräftebefragung zum Lernen mit digitalen Medien im Vergleich. In S.G. Huber, C. Helm, & N. Schneider (Hrsg.), *COVID-19 und Bildung – Studien und Perspektive*n. Münster, New York: Waxmann. https://doi.org/10.31244/9783830996361

Zusammenfassung

Der durch die Corona-Pandemie forcierte Distanzunterricht stellte Lehrkräfte vor neuartige Herausforderungen, um Schülerinnen und Schüler weiterhin zu unterrichten. Spontan nahmen digitale Medien eine zentrale Rolle im Unterrichtsgeschehen ein. In zwei online Umfragen wurden im Frühjahr 2020 (etwa zur Zeit der Schulschließungen) 2810 Lehrkräfte und vor den Osterferien 2021 (während und nach erneuter großflächiger Schulschließungen) insgesamt 1774 Lehrkräfte deutschlandweit über ihre Erfahrungen und Eindrücke zum Lernen und Lehren mit digitalen Medien während des Distanzunterrichts befragt. Im Fokus der Untersuchungen standen Fragen dazu, wie es um die technische Ausstattung an den Schulen für den Distanzunterricht stand, wie mit Schülerinnen und Schülern kommuniziert wurde, welche Einstellungen Lehrkräfte gegenüber dem Einsatz digitaler Medien hatten, wie sie ihre Fähigkeiten zum Unterrichten mit digitalen Medien einschätzten und wie die Lehrkräfte den Unterricht gestalteten. Dieser Beitrag geht vor dem Hintergrund dieser Studien der Frage nach, wie sich der Distanzunterricht zwischen 2020 und 2021 aus der Perspektive der Lehrkräfte entwickelt hat. Vergleiche zwischen den Erhebungszeitpunkten zeigen, dass Unterricht 2021 vermehrt in digitaler Form über Konferenzprogramme und Lernplattformen stattgefunden hat. Auch wurde eine Verbesserung der technischen Ausstattung der Schülerinnen und Schüler beobachtet, welche zudem vermehrt Ausleihmöglichkeiten für Geräte wahrnehmen konnten. Mit dem verstärkten Nutzen von digitalen Medien zur Unterrichtsgestaltung ließ sich auch ein Zuwachs in der Selbstwirksamkeit von Lehrkräften im Umgang mit digitalen Medien feststellen. Lehrkräfte suchten und modifizierten zudem vermehrt Unterrichtsmaterialien aus Internetquellen. Probleme mit der Technik, Ausstattung und Internetverbindung bestanden jedoch nach wie vor und schränkten Lehrkräfte im virtuellen Unterricht ein. Ursachen für die Veränderungen sowie Ausblicke für die weitere Digitalisierung im Unterricht werden diskutiert.

1. Einleitung: Unterricht während der Corona-Pandemie

Die schulische Lehr-Lernsituation in Form einer Präsenz von Lehrkräften sowie Schülerinnen und Schülern war bislang Standard des Unterrichtens in Deutschland und vielen anderen Ländern. Die Corona-Pandemie führte im März 2020 zu einer rapiden und bis dahin für die beteiligten Akteure nicht gekannten Umstellung dieses Unterrichts, weg vom Präsenzunterricht hin zu Distanzunterricht, Wechselunterricht mit Teilen der Klassen oder hybriden Formen. Dies ging mit enormen Herausforderungen für die Beteiligten einher: für die Lehrkräfte, die den Unterricht unvermittelt umgestalten mussten, ohne Vorerfahrungen in diesem Kontext zu haben, für die Schülerinnen und Schüler, die ihre Lernprozesse zu Hause in neuer Form organisieren mussten, und auch für Eltern, die neben ihren beruflichen Anforderungen das Lernen der Kinder begleiten sollten (Fickermann & Edelstein, 2020; Huber et al., 2020; Steinmayr et al., 2021; Voss & Wittwer, 2020; Wößmann et al., 2020). Die Corona-Pandemie brachte neben den allgemeinen und gesellschaftlichen Auswirkungen auch mit sich, dass Lehr-Lernprozesse umgestaltet werden mussten, um Bildungsprozesse der Kinder und Jugendlichen aufrechterhalten zu können. Dabei spielte die Digitalisierung eine zentrale Rolle, die beispielsweise mit Möglichkeiten der digitalen Kommunikation und Kollaboration sowie synchronen (z.B. gleichzeitige Lernsituationen per Videokonferenz) und asynchronen Formaten (z.B. zeitlich flexibilisierte Lernsituationen mithilfe von Lernmanagementsystemen) zahlreiche Optionen für die Gestaltung der Lernprozesse bietet.

Die Digitalisierung in Schulen ist in Deutschland ein Feld, dem in den vergangenen Jahren mehr und mehr Aufmerksamkeit zuteilwurde. Dies wurde nicht zuletzt durch empirische Befunde begründet, die Deutschland Nachholbedarfe attestierten. Erforderlich sind eine verstärkte Implementation digitaler Medien und die Förderung der Kompetenzen von Schülerinnen und Schülern im eigenständigen sowie kritisch-reflektierten Umgang mit digitalen Medien, gleichzeitig wurde auch die Notwendigkeit einer flächendeckenden medienpädagogischen Aus- und Fortbildung von Lehrkräften betont (Eickelmann et al., 2019; McElvany, 2018; vbw, 2018). Bildungspolitisch mündeten diese Befunde in Vorgaben und Strategiepapiere zur gezielten Weiterentwicklung dieses Bereichs, der für gesellschaftliche, berufliche und individuelle Teilhabe sowie lebenslanges Lernen als relevant erachtet wird. Sowohl bundesweite Rahmungen (BMBF, 2016; KMK, 2016) als auch Entwicklungen in den einzelnen Ländern der Bundesrepublik Deutschland stützen die Weiterentwicklung im Kontext der Digitalisierung in Schule und Unterricht. Finanziell ist diese Weiterentwicklung mit Maßnahmen in den Bundesländern begleitet worden und seitens des Bundes wurde mit

dem DigitalPakt Schule in die digitale Ausstattung der Schulen investiert – bei gleichzeitiger Verpflichtung der Länder, die Fortbildung der Lehrkräfte im Kontext der Digitalisierung massiv auszubauen und flächendeckend zugänglich zu machen (BMBF, 2019).

Vor diesem Hintergrund stellte sich die Frage, wie Lehrkräfte während der Corona-Pandemie Unterricht gestalteten und die Digitalisierung einbezogen haben. Insbesondere angesichts der fortdauernden Herausforderungen und Einschränkungen des regulären Präsenzunterrichts war von Interesse, welche Veränderungen im Kontext digital gestützter Lehr-Lernsituationen über die Zeit hinweg beobachtet werden konnten. Zur Untersuchung dieser zentralen Fragen zum Unterricht während der Corona-Pandemie führte ein Forschungsteam des Instituts für Schulentwicklungsforschung (IFS) an der Technischen Universität Dortmund zwei bundesweite Befragungen mit Lehrkräften allgemeinbildender Schulen aller Schulformen in Deutschland durch – eine zum Zeitpunkt der ersten bundesweiten Schulschließungen im April 2020 und eine zweite ein knappes Jahr später während erneuter coronabedingter Schulschließungen in Deutschland. Veränderungen im professionellen Handeln der Lehrkräfte bezüglich des digitalisierten Unterrichtens konnten auf dieser Basis erforscht werden.

Im Folgenden werden nach einem Überblick zu der Ausgangslage für digital unterstütztes Lehren und Lernen während der Pandemie (Abschnitt 2) zunächst die Durchführung und die Teilnehmenden der Studie vorgestellt (Abschnitt 3). Anschließend werden die Befunde zu den technischen Grundvoraussetzungen und Rahmenbedingungen des Lehrens und Lernens mit digitalen Medien während der Pandemie beschrieben (Abschnitt 4.1), bevor die Befunde zur Kommunikation zwischen Lehrenden und Lernenden (Abschnitt 4.2) dargelegt werden. Mit Fokus auf Veränderungen für die Lehrkräfte werden Unterschiede in den Einstellungen (Abschnitt 4.3) sowie ihren selbst eingeschätzten Kompetenzen (Abschnitt 4.4) zum Einsatz digitaler Medien in Lehr-Lernprozessen vorgestellt. Schließlich wird der Frage nachgegangen, inwiefern Lehrkräfte digitale Medien für das Lehren und Lernen im Verlauf der Pandemie einsetzen (Abschnitt 4.5). Der Beitrag schließt mit einer Zusammenfassung und Diskussion der Befunde (Abschnitt 5).

2. Design

2.1 Hintergrund und theoretischer Rahmen: Lehren und Lernen mit digitalen Medien während der Pandemie

Die Corona-Krise hat in Deutschland bereits zweimal zu bundesweiten Schulschließungen geführt, die zur Eindämmung des Infektionsgeschehens beitragen sollten. Für die Beteiligten stellte dies insbesondere zu Beginn der Pandemie im Frühjahr 2020 enorme Herausforderungen dar, da ein Wechsel in den Distanzunterricht kurzfristig und ohne Vorbereitung oder Vorerfahrung erfolgen musste. Lehrkräfte standen somit vor der Aufgabe, diese Situation, die rückblickend in der Folge dieser Schulschließungen vielfach zu weiteren Abweichungen des Regelunterrichts in Präsenz geführt hat, so zu gestalten, dass die Lernprozesse der Kinder und Jugendlichen aufrechterhalten werden konnten. Dies ging vielfach mit dem Einsatz digitaler Medien einher, die für vielfältige Zwecke genutzt wurden wie beispielsweise für die Kommunikation, Koordination oder Gestaltung und Übermittlung von Aufgaben. Die Möglichkeiten, die digitale Medien in diesem Kontext offerieren, waren zwar nicht neu, jedoch vielfach kaum und vor allem nicht ausschließlich im Distanzunterricht erprobt.

Um die Phase der ersten Schulschließungen und die Nutzung digitaler Medien rahmen zu können, muss die Ausgangslage, die in Bezug auf die Digitalisierung in den Schulen in Deutschland vor der Pandemie vorzufinden war, betrachtet werden. Empirische Befunde zeigten wiederholt, dass die technische Ausstattung der Schulen in Deutschland im internationalen Vergleich wenig ausgebaut war. Auch innerhalb Deutschlands waren deutliche Unterschiede vorzufinden, sowohl zwischen unterschiedlichen Schulformen als auch zwischen Bundesländern (Eickelmann et al., 2019; Lorenz & Endberg, 2017). Die IT-Ausstattung wird als notwendige Voraussetzung für den Einsatz digitaler Medien für unterrichtliche Zwecke angesehen, sie ist jedoch noch nicht hinreichend, um eine lernförderliche Integration digitaler Medien im Unterricht sicherzustellen. Zu Pandemiezeiten hat sich dieser Stellenwert allerdings gewandelt, da Unterrichtsorganisation und Kommunikation zwischen Lehrenden und Lernenden während des Distanzunterrichts in erheblichem Maße auf diese Grundvoraussetzungen angewiesen sind. Im Zuge der Corona-Pandemie wurde daher im Rahmen des DigitalPakts eine flexible und schnelle Hilfe vereinbart, indem die zugesagten Mittel von fünf Milliarden Euro um weitere eineinhalb Milliarden für Voraussetzungen zur Erstellung digitaler Inhalte, ausleihbare mobile Endgeräte

für Schülerinnen und Schüler oder Leihgeräte für Lehrkräfte ergänzt wurden[3]. Die Länder bzw. Schulträger tragen nochmals zehn Prozent von dieser Summe bei. Damit sollen die Voraussetzungen für eine Bildung in der digitalen Welt nachhaltig verbessert werden. Somit stellt sich die Frage, ob zum zweiten Zeitpunkt der bundesweiten Schulschließungen im Jahr 2021 in den Schulen ein Unterschied in den technischen Grundvoraussetzungen und Rahmenbedingungen des digital gestützten Unterrichts zu verzeichnen ist.

Die Kommunikation zwischen Lehrkräften und ihren Schülerinnen und Schülern kann während des Distanzunterrichts als Herzstück der Interaktion und des Aufrechterhaltens von Lehr-Lernprozessen angesehen werden. Eine wesentliche Grundlage dafür stellen digitale Medien dar, mithilfe derer die Kommunikation und Vermittlung auf Seiten der Lehrenden wie auch der Lernenden erfolgen kann (z.B. Köller et al., 2020). Verschiedene (analoge und digitale) Kommunikationsmittel stehen hier zur Verfügung, die synchrone und asynchrone Kommunikation mit einer Lerngruppe oder einzelnen Schülerinnen und Schülern ermöglichen. Zu Beginn der Pandemie hat sich gezeigt, dass die Kommunikation insbesondere via E-Mails, Briefen und Telefon aufrechterhalten wurde (Anger et al., 2020; BiSE, 2020; Lorenz et al., 2020; Sander et al., 2020). Hier stellt sich die Frage, ob im Laufe der Pandemie Kommunikationswege gleichgeblieben sind, oder ob andere Kommunikationswege intensiver genutzt werden.

Lehrkräfte gelten als „keystone species“ für die Integration digitaler Medien in schulisches Lernen (Davis et al., 2013, S. 439). Daher sind Einstellungen und Kompetenzen von Lehrkräften zum Einsatz digitaler Medien in Lehr-Lernprozessen von äußerster Relevanz. Verschiedene nationale wie internationale Studien zeigten, dass die Einstellungen der Lehrkräfte zum Einsatz von digitalen Medien im Unterricht eine der größten Barrieren oder Katalysatoren für eine nachhaltige Integration und Nutzung zur Unterstützung des Lernens in Schulen sind (Drent & Meelissen, 2008; Eickelmann & Vennemann, 2017; Lorenz et al., 2019; Tondeur et al., 2016). Demgemäß zeigten Forschungsergebnisse, dass die Einstellungen und Überzeugungen der Lehrkräfte gegenüber dem Einsatz digitaler Medien signifikante Prädiktoren der Häufigkeit und Qualität der Nutzung digitaler Medien darstellen (Celik & Yesilyurt, 2013; Ertmer, 2005; Scherer et al., 2020; Siyam, 2019). Daher ist es von besonderer Relevanz, dass sich Lehrkräfte im Umgang mit digitalen Medien (Knezek & Christensen,

[3] https://www.digitalpaktschule.de/de/der-digitalpakt-und-die-corona-krise-1784.html [Zugriff am 23.04.2021]

2018), denen insbesondere während der Corona-Pandemie eine besondere Bedeutung zukommt, wohlfühlen.

Die Qualifizierung der Lehrkräfte im Kontext der Digitalisierung ist eine zentrale Aufgabe der Lehrerbildung (KMK, 2016; van Ackeren et al., 2019). Lehrkräfte müssen die Strategien für eine effektive Integration von digitaler Technologie in das Lehren und Lernen verstehen, um diese lernförderlich nutzen zu können (Knezek & Christensen, 2018). Mit dem vielfach herangezogenen TPACK-Modell (als Akronym für technological pedagogical and content knowledge) haben Mishra und Koehler (2006) einen theoretischen Ansatz entwickelt, mit dem die Wissensdomänen des technologischen, pädagogischen und (fach-)inhaltlichen Wissens kombiniert betrachtet werden können. Mit verschiedenen Strategien, wie z.B. dem Beobachten, Anwenden oder Reflektieren über den Einsatz digitaler Medien im Unterricht, können Lehrkräfte sich dieses Wissen aneignen (Tondeur et al., 2019). Untersuchungen konnten die Relevanz des Wissens von Lehrkräften in diesem Bereich für die unterrichtliche Nutzung digitaler Medien herausstellen (Seufert et al., 2021). Interessant ist nun die Frage, ob im Verlauf der Pandemie aufgrund der intensivierten Erfahrungen mit digitalen Medien und digital gestütztem Lehren und Lernen Unterschiede identifiziert werden können: Haben sich – im Vergleich zum Beginn – nach einem Jahr der Pandemie die Einstellungen oder die Selbsteinschätzung der Lehrkräfte zu ihren Kompetenzen im Kontext des digitalen Unterrichts verändert?

Schließlich stellten das Lehren und Lernen mit digitalen Medien selbst den Kern des Geschehens dar. Der Distanzunterricht kann von Potenzialen des Medieneinsatzes profitieren, beispielsweise durch direkten Kontakt zwischen Lehrkräften und Schülerinnen und Schülern für die Besprechung von Inhalten, die Nutzung von Apps zur Förderung fachlicher Leistungen oder die Erstellung und bedarfsgerechte Modifizierung digitaler Unterrichtsmaterialien. Veränderungen im Laufe der Pandemie sind auch hier denkbar.

2.2 Zentrale Forschungsfragen

Vor diesem Hintergrund stellte sich die Frage, wie sich das Lernen mit digitalen Medien im Laufe der Pandemie verändert hat. Diesbezüglich wurde mit dem vorliegenden Beitrag auf Grundlage unserer Forschung den folgenden Fragen nachgegangen:

1. Wie waren die technischen Grundvoraussetzungen für das Lehren und Lernen im Distanzunterricht zu Beginn der Pandemie und zur Zeit der zweiten bundesweiten Schulschließungen knapp ein Jahr später?

2. Wie fand die Kommunikation zwischen Lehrenden und Lernenden in diesen beiden Phasen der Schulschließungen statt?
3. Lassen sich Unterschiede in den Einstellungen der Lehrkräfte zum Einsatz digitaler Medien im Unterricht zu den beiden Zeitpunkten ausmachen?
4. Gibt es Unterschiede in der Selbsteinschätzung der Kompetenzen der Lehrkräfte zum Einsatz digitaler Medien in Lernsituationen zu Beginn und nach einem Jahr der Pandemie?
5. Wie fand Lehren und Lernen digital gestützt während der beiden Phasen der Schulschließungen statt?

2.3 Methode – Durchführung der Studie

Zur Untersuchung von Fragestellungen zum Unterricht während der Corona-Pandemie wurde im Frühjahr 2020 eine bundesweite Onlinebefragung von Lehrkräften durchgeführt. Die webbasierte Befragung mit LimeSurvey erfolgte im Zeitraum von Mitte April bis Ende Mai 2020. Lehrkräfte allgemeinbildender Schulen in allen 16 Bundesländern wurden über vielfältige Kanäle (u.a. Facebook-Werbung, Emails, Twitter) auf die Befragung aufmerksam gemacht und zur Teilnahme eingeladen. Die Teilnahme an dieser Befragung erfolgte anonym und freiwillig. Die Beantwortung nahm etwa 15 Minuten in Anspruch und umfasste neben Fragen zur Soziodemografie unter anderem auch Fragen zum Wohlbefinden der Lehrkräfte, zum Unterricht sowie zu Einstellungen und Kompetenzen hinsichtlich der Digitalisierung. Soziodemografische Fragen umfassten das Bundesland, die Schulform, das Alter, das Geschlecht und die Erfahrung der Lehrkräfte im Schuldienst. Weiterhin wurden Lehrkräfte mit fünf Einzelitems nach den technischen Gegebenheiten im Distanzunterricht gefragt. Lehrkräfte konnten ihre Zustimmung mit Aussagen wie z.B. „Während den Wochen der Schulschließungen aufgrund der Corona-Pandemie gab es viele Technikprobleme (z.B. mit Hardware, Internetverbindung, Lernplattform)“ entweder mit „stimme zu“, „stimme eher zu“, „stimme eher nicht zu“, oder mit „stimme nicht zu“ angeben. Auch wurden Lehrkräfte gefragt, mit wie vielen ihrer Schülerinnen und Schüler sie verschiedene Kontaktmedien wie z.B. Briefe, Emails oder Lernplattformen nutzten. Dabei konnten Lehrkräfte zwischen den Antwortmöglichkeiten „mit allen“, „mit der Mehrheit“, „mit einigen“ oder „mit keinem“ auswählen. Einstellungen der Lehrkräfte zum Einsatz von digitalen Medien im Unterricht wurde mit fünf Items gemessen. Wieder konnten Lehrkräfte ihre Zustimmung zu Aussagen wie z.B. „Der Einsatz digitaler Medien erhöht die Motivation der Schülerinnen und Schüler“

auf einer vierstufigen Likert-Skala angeben, die von „stimme zu“ bis „stimme nicht zu“ reichte. Die Selbsteinschätzung von Lehrkräften zu ihrer Kompetenz mit digitalen Medien wurde mithilfe einer Skala zum technological pedagogical and content knowledge (TPACK) abgefragt (Schmidt et al., 2009). Diese besteht aus fünf Aussagen (z.B. „Ich kann digitale Medien auswählen, mit denen sich die Fachinhalte im Unterricht besser vermitteln lassen.“), die Lehrkräfte mit „stimme zu“, „stimme eher zu“, „weder Zustimmung noch Ablehnung“, „stimme eher nicht zu“ und „stimme nicht zu“ beantworten konnten. Schließlich wurden mehrere Fragen dazu gestellt, wie Lehrkräfte den Distanzunterricht gestalten. Auch hier wurde eine vierstufige Likert-Skala eingesetzt die von „stimme zu“ bis „stimme nicht zu“ reichte. Fragen umfassten beispielsweise die Nutzung von Lern-Apps, von virtuellem Unterricht und die Suche und Modifikation von neuen Unterrichtsaufgaben für die Schülerinnen und Schüler.

2.4 Stichprobe – Beschreibung der Teilnehmenden

Für die erste Befragung 2020 liegen vollständige Angaben von bundesweit 2.810 Lehrkräften[4] aus allgemeinbildenden Schulen vor. Der größte Teil der Teilnehmenden war weiblich (82.0 % weiblich; 18.0 % männlich) und im Durchschnitt 40 Jahre alt (M = 40.3; SD = 9.71). Die Befragten stammten aus allen 16 Bundesländern der Bundesrepublik Deutschland, wobei Lehrkräfte aus Bayern (20.8 %) und Nordrhein-Westfalen (29.8 %) verstärkt repräsentiert waren. Die Erfahrung im Schuldienst betrug im Durchschnitt 12.9 Jahre, wobei die berufliche Erfahrung als Lehrkraft individuell stark variierte (SD = 9.20).
Nach anhaltenden Auswirkungen der Pandemie auf die Schulen und auf den Unterricht sowie aufgrund der erneuten Schulschließungen zu Beginn des Jahres 2021 wurde die Onlinebefragung wiederholt. Inhaltlich umfasste die zweite Befragung die gleichen Aspekte, sodass Veränderungen im Laufe der Corona-Pandemie untersucht werden können. Die zweite Erhebung wurde um differenziertere Fragen zum Einsatz von Lernmanagementsystemen ergänzt. Der Erhebungszeitraum erstreckte sich hier von Ende Januar bis Mitte März 2021. An der zweiten Onlinebefragung, die auf die gleiche Weise administriert und verbreitet wurde, nahmen 1774[5] Lehrkräfte teil und beantworteten den Fragebogen vollständig. Auch unter den Teilnehmenden der zweiten Befragung war der Anteil weiblicher Lehrkräfte größer (84.2 %

[4] Die Angaben von 276 Schulleitungen wurden hier nicht berücksichtigt.

[5] Die Angaben von 154 Schulleitungen wurden hier nicht berücksichtigt.

weiblich; 15.8 % männlich) und die Befragten waren im Durchschnitt ca. 43 Jahre alt (M = 43.3; SD = 9.67). An der Befragung nahmen Lehrkräfte aus allen 16 Bundesländern teil, wobei Lehrkräfte aus Nordrhein-Westfalen (31.0 %) verstärkt repräsentiert waren. Im Mittel verfügten die teilnehmenden Lehrkräfte über 15.4 Jahre Erfahrung im Schuldienst, die Dauer der beruflichen Erfahrung variierte individuell stark (SD = 9.62).

Die vertretenen Schulformen zu den zwei Erhebungszeitpunkten sind in *Abbildung 1* dargestellt. Zu erwähnen ist dabei, dass in der Erhebung 2020 die Kategorie *Berufskolleg oder vergleichbar* nicht differenziert erfasst wurde. Die prozentuale Verteilung der Teilnehmenden nach Schulform unterschied sich signifikant zwischen den Erhebungszeitpunkten.

Abbildung 1

Prozentuale Verteilung der Lehrkräfte nach Schulform

Die Daten wurden mit dem Programm SPSS 27 ausgewertet. Für die im Folgenden berichteten Analysen wurde eine Gewichtung genutzt, um eine Anpassung der realisierten Stichproben an die Merkmalsverteilung von Lehrkräften in Deutschland unter Berücksichtigung des Geschlechts, des Alters, der Schulform und des Bundeslands zu erreichen. In der Darstellung der Ergebnisse werden nur gültige Prozente berichtet. Zudem werden Unterschiede zwischen Subgruppen nur dann berichtet, wenn diese statistisch signifikant waren.

3. Befunde zu Veränderungen des Unterrichts während der Pandemie

Die Corona-Pandemie hatte weitreichende Konsequenzen für Lehrkräfte, Schülerinnen und Schüler und die Gestaltung von Lehr-Lernprozessen. Der Digitalisierung kam dabei eine wesentliche Funktion zu, nur mit ihr konnten diese Prozesse im Wesentlichen aufrechterhalten werden. Im folgenden Ergebnisteil wird daher verstärkt auf den Einsatz digitaler Medien geblickt und auch der Frage nachgegangen, welche Unterschiede nach einem Jahr der Pandemie in den Schulen deutlich wurden.

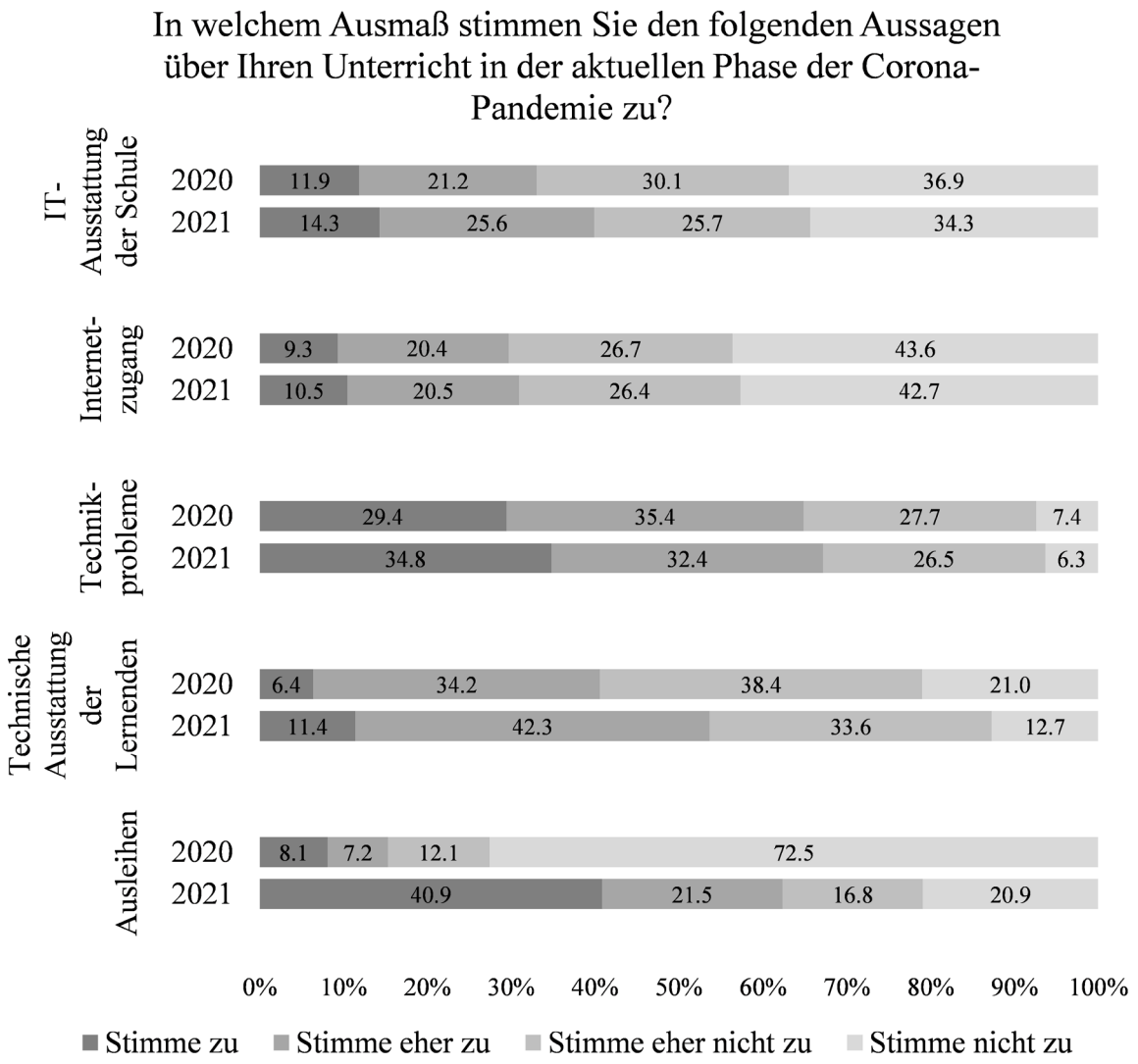
3.1 Technische Grundvoraussetzungen und Rahmenbedingungen des Lehrens und Lernens mit digitalen Medien

Das Lernen während der Pandemie entfernte sich nicht nur in weiten Teilen vom Lernort Schule, es bezog auch in erheblichem Maße die Möglichkeiten der Digitalisierung mit ein. Eine Grundvoraussetzung dafür stellte die technische Ausstattung für das Lehren und Lernen dar, aber auch passgenaue Rahmenbedingungen im Hinblick auf die schulinternen Prozesse waren notwendig. Die Lehrkräfte wurden daher im Fragebogen gebeten, die technischen Voraussetzungen des Lehrens und Lernens mit digitalen Medien einzuschätzen. Dazu sollte anhand einer vierstufigen Antwortskala (1 = stimme zu; 4 = stimme nicht zu) zum einen die ausreichende Funktionalität und zum anderen bewertet werden, inwiefern die technische Ausstattung der Lernenden auf Schülerseite sichergestellt war. Darüber hinaus wurden Lehrkräfte um ihre Einschätzung der schulinternen Rahmenbedingungen zur Gestaltung digitalisierungsgestützter Unterrichtseinheiten gebeten. Die Befunde werden im Vergleich zwischen den beiden Erhebungszeitpunkten 2020 und 2021 berichtet, sodass Veränderungen im Verlauf der Pandemie betrachtet werden können.

Die Ergebnisse verdeutlichen, dass die Lehrkräfte insgesamt eher nicht mit den technischen Voraussetzungen zufrieden waren, jedoch im Vergleich der beiden Erhebungszeitpunkte positive Tendenzen hinsichtlich der Weiterentwicklung der technischen Rahmenbedingungen aufgezeigt werden konnten. So stimmten die Lehrkräfte der Aussage, dass eine ausreichende IT-Ausstattung (z.B. Computer, Software) vorhanden war, insgesamt eher nicht zu, wobei der Anteil jener Lehrkräfte, die dieser Aussage zustimmten oder eher zustimmten, zum zweiten Erhebungszeitpunkt höher war und im Mittel ein signifikanter Unterschied vorliegt (2020: $M = 2.92$; $SD = 1.02$; 2021: $M = 2.80$; $SD = 1.07$). Die Anteile der Lehrkräfte, die der Aussage zustimmten, können Abbildung 2 entnommen werden.

Abbildung 2

Zustimmung der Lehrkräfte zu Aspekten der technischen Ausstattung in Prozent



Notiz. IT-Ausstattung der Schule: "Es ist eine ausreichende IT-Ausstattung vorhanden (z. B. Computer, Software)."; Internetzugang: "Der Internetzugang ist ausreichend (z. B. Geschwindigkeit und Stabilität der Verbindung)."; Technikprobleme: "In diesen Woche gab es viele Technikprobleme (z.B. mit Hardware, Internetverbindung, Lernplattform)."; Technische Ausstattung der Lernenden: "Die technische Ausstattung der Schülerinnen und Schüler war ausreichend, um den Unterricht digital gestützt fortzuführen."; Ausleihen: "Bei

unzureichender Ausstattung konnten die Schülerinnen und Schüler digitale Medien von der Schule ausleihen, um zu Hause angemessen arbeiten zu können."

Hinsichtlich des Internetzugangs (z.B. Geschwindigkeit und Stabilität der Verbindung) waren die Lehrkräfte wenig zufrieden und zwischen den beiden Erhebungszeitpunkten ist diesbezüglich kein Unterschied festzustellen (2020: $M = 3.05$; $SD = 1.01$; 2021: $M = 3.01$; $SD = 1.03$). Darüber hinaus gaben die Lehrkräfte insgesamt eher an, dass es viele Technikprobleme gab, wobei der Anteil zum zweiten Erhebungszeitpunkt 2021 im Mittel signifikant höher war (2020: $M = 2.13$; $SD = 0.92$; 2021: $M = 2.04$; $SD = 0.93$). Deutlichere Fortschritte konnten im Bereich der schülerseitigen Ausstattung verzeichnet werden. Zum einen stimmten die Lehrkräfte der Aussage, dass die technische Ausstattung der Schülerinnen und Schüler für die Fortführung eines digital gestützten Unterrichts ausreichend war, im Jahr 2020 eher nicht zu und im Jahr 2021 im Mittel eher zu (2020: $M = 2.74$; $SD = 0.86$; 2021: $M = 2.48$; $SD = 0.86$). Zum anderen hat sich die marginal vorhandene Möglichkeit für Schülerinnen und Schüler, digitale Medien von der Schule ausleihen zu können, vom Jahr 2020 ($M = 3.49$; $SD = 0.94$) bis zum zweiten Erhebungszeitpunkt 2021 deutlich verbessert ($M = 2.18$; $SD = 1.18$).

Die weiteren schulinternen Rahmenbedingungen für die Gestaltung digital angelegter Unterrichtseinheiten wurden von den Lehrkräften eher kritisch eingeschätzt und der Vergleich zwischen den Erhebungszeitpunkten zeigte eher geringe Veränderungen. So schätzten die Lehrkräfte die zur Verfügung stehende Zeit zur Vorbereitung computergestützter Unterrichtsstunden kritischer ein und stimmten im Mittel zum zweiten Erhebungszeitpunkt eher weniger zu, dass die Vorbereitungszeit genügte (Ergebnisse 2020: $M = 3.18$; $SD = 0.86$; stimme zu: 5.2 %, stimme eher zu: 13.6 %, stimme eher nicht zu: 38.8 %, stimme nicht zu: 42.4 %; Ergebnisse 2021: $M = 3.30$; $SD = 0.80$; stimme zu: 3.3 %, stimme eher zu: 11.6 %, stimme eher nicht zu: 36.5 %, stimme nicht zu: 48.6 %). Der Aussage, dass die Schulleitung großen Wert auf den Einsatz digitaler Medien im Unterricht legt, stimmten die Lehrkräfte im Durchschnitt eher zu, wobei keine statistisch signifikanten Unterschiede zwischen den Erhebungszeitpunkten vorlagen (Ergebnisse 2020: $M = 2.42$; $SD = 0.99$; stimme zu: 19.9 %, stimme eher zu: 34.9 %, stimme eher nicht zu: 28.1 %, stimme nicht zu: 17.1 %; Ergebnisse 2021: $M = 2.38$; $SD = 1.02$; stimme zu: 22.6 %, stimme eher zu: 34.9 %, stimme eher nicht zu: 24.6 %, stimme nicht zu: 17.8 %). Die Lehrkräfte waren zu beiden Zeitpunkten eher nicht der Meinung, dass es an ihrer Schule genügend Beispielmaterial zu digital gestütztem

Unterricht gibt, sodass hier im Laufe des Jahres während der Pandemie keine Veränderungen ersichtlich wurden (Ergebnisse 2020: $M = 3.05$; $SD = 0.87$; stimme zu: 4.9 %, stimme eher zu: 20.3 %, stimme eher nicht zu: 39.3 %, stimme nicht zu: 35.5 %; Ergebnisse 2021: $M = 3.08$; $SD = 0.86$; stimme zu: 4.1 %, stimme eher zu: 21.3 %, stimme eher nicht zu: 37.4 %, stimme nicht zu: 37.1 %). Der Aussage, dass interne Workshops zu digital gestütztem Unterricht regelmäßig im Kollegium organisiert werden, stimmten die Lehrkräfte im Mittel ebenfalls eher nicht zu, wobei im Vergleich der Jahre eine signifikante positive Tendenz verzeichnet werden kann (Ergebnisse 2020: $M = 3.21$; $SD = 0.94$; stimme zu: 6.3 %, stimme eher zu: 16.5 %, stimme eher nicht zu: 27.1 %, stimme nicht zu: 40.1 %; Ergebnisse 2021: $M = 2.95$; $SD = 1.01$; stimme zu: 10.6 %, stimme eher zu: 22.0 %, stimme eher nicht zu: 29.4 %, stimme nicht zu: 38.0 %).

Insgesamt zeigte sich somit, dass die technischen Voraussetzungen und die Rahmenbedingungen des Lehrens und Lernens mit digitalen Medien während der Corona-Pandemie eher kritisch von den Lehrkräften eingeschätzt wurden. Trotz positiver Entwicklungstendenzen für die Lehrkräfte sind insgesamt geringe Veränderungen festzustellen, sodass Unterricht weiterhin unter eher erschwerten Bedingungen gestaltet werden musste. Ein deutlicher Fortschritt konnte hingegen im Bereich der Ausstattung der Schülerinnen und Schüler verzeichnet werden, insbesondere im Hinblick auf Ausleihmöglichkeiten digitaler Geräte von der Schule.

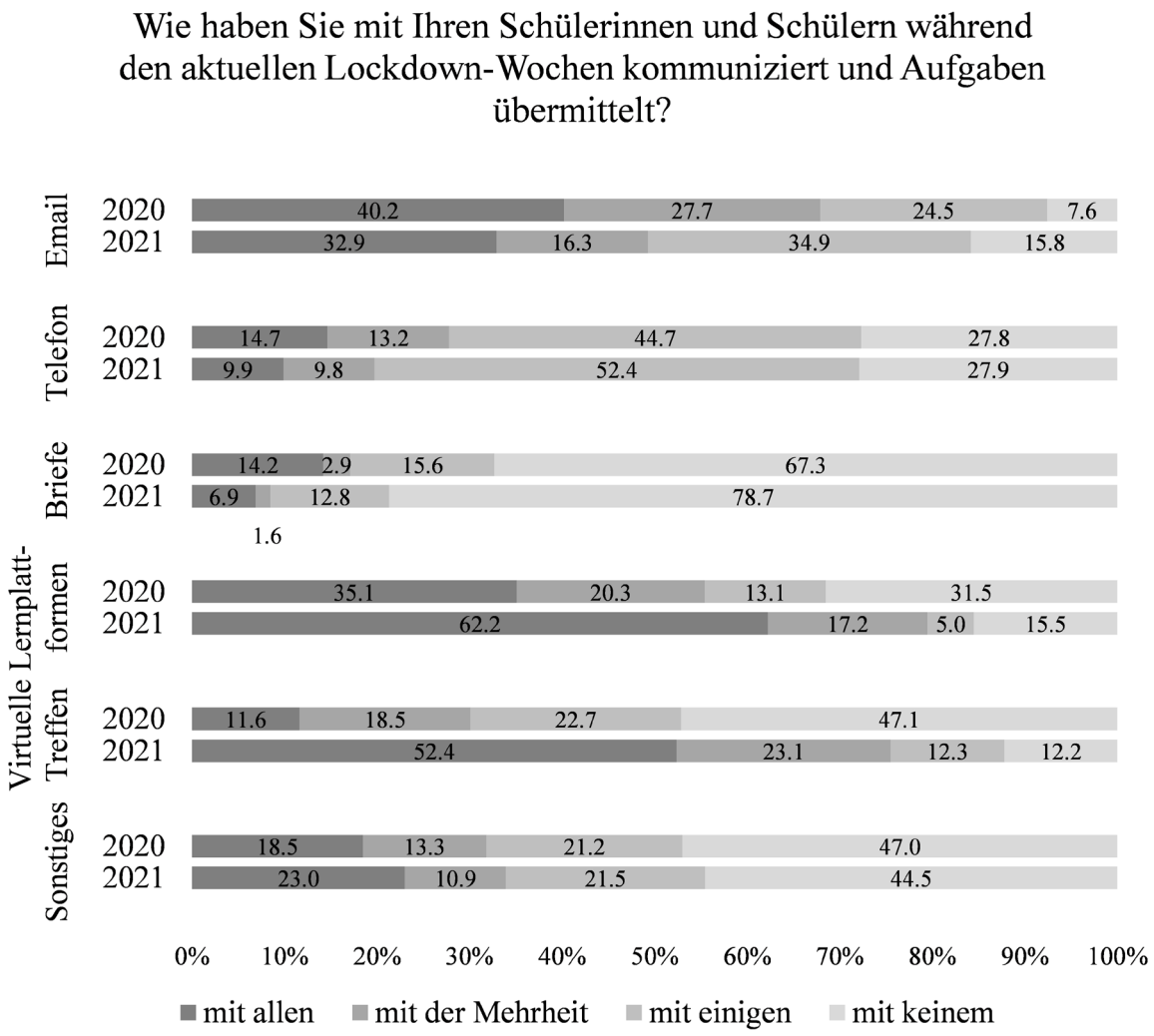
3.2 Kommunikation zwischen Lehrenden und Lernenden

Die Wochen des Lockdowns erforderten einen Distanzunterricht, in dem die reguläre Interaktion des Präsenzunterrichts vollkommen ausgesetzt werden musste. Verschiedene Kontaktmöglichkeiten konnten in dieser Zeit die Kommunikation zwischen Lehrenden und Lernenden weiterhin ermöglichen und den Austausch von Materialien gestatten. Diesbezüglich wurden die Lehrkräfte gebeten anzugeben, wie sie während der Wochen des Lockdowns mit den Schülerinnen und Schülern kommuniziert haben und ob diese Kommunikation mit der gesamten Klasse, mit Gruppen bzw. einzelnen Schülerinnen und Schülern über das jeweils angegebene Medium stattfand. Die Antworten wurden mithilfe einer vierstufigen Skala erfasst (1 = mit allen Lernenden, 2 = mit der Mehrheit, 3 = mit einigen, 4 = mit keinem). Als Kontaktmedien wurden E-Mails, Telefon, Briefe, Lernplattformen oder virtuelle Treffen sowie eine Kategorie *Sonstiges* erfasst. Abbildung 3 zeigt die Anteile der

Lehrpersonen, die entsprechenden Kontaktmedien mit den Schülerinnen und Schülern in den Wochen des Lockdowns nutzten.

Abbildung 3

Verteilung Angaben zu Kommunikationswegen zwischen Lehrkräften und Schülerinnen und Schülern in Prozent



Hinsichtlich der Kommunikationsmedien wird ersichtlich, dass E-Mails und Lernplattformen eher intensiv genutzt wurden. Telefon und Briefe wurden hingegen weniger als Kommunikationsmedium genutzt. Der Vergleich zwischen den Jahren zeigte hinsichtlich aller Kommunikationswege signifikante Veränderungen und ist besonders im Hinblick auf virtuelle Treffen auffällig. Zum zweiten Erhebungszeitpunkt 2021 gaben mehr als die Hälfte der Lehrkräfte an, mit allen Schülerinnen und Schülern virtuelle Treffen abzuhalten, während der Wert 2020 noch bei 11.6 Prozent lag. In ähnlicher Weise lässt sich auch mit Blick auf den Einsatz von Lernplattformen ein intensiverer Austausch feststellen: Während 2020 noch 35.1

Prozent der Lehrpersonen angaben, mit allen Lernenden per Lernplattformen zu kommunizieren und Aufgaben zu übermitteln, lag der Anteil 2021 mit 62.2 Prozent deutlich höher. Die Kategorie *Sonstiges* wurde als offenes Antwortformat administriert. Die Auswertung der Angaben zeigte, dass über die erfassten Kommunikationswege hinaus Aufgabenmaterial von den Schulen ausgegeben wurde, Materialien per Brief versendet wurden oder Lehrkräfte die Übergabe teils selbst zum Briefkasten der Schülerfamilien oder per direkter Übergabe organisiert haben.

Insgesamt zeigte sich im Verlauf der Pandemie eine Abnahme individueller Kommunikationsformen wie das Übermitteln von Briefen oder E-Mails und eine Zunahme der Gruppenkommunikation via Lernplattformen und virtueller Treffen. Insbesondere für die beiden letztgenannten zeigte sich im Laufe der Pandemie eine deutlich höhere Relevanz.

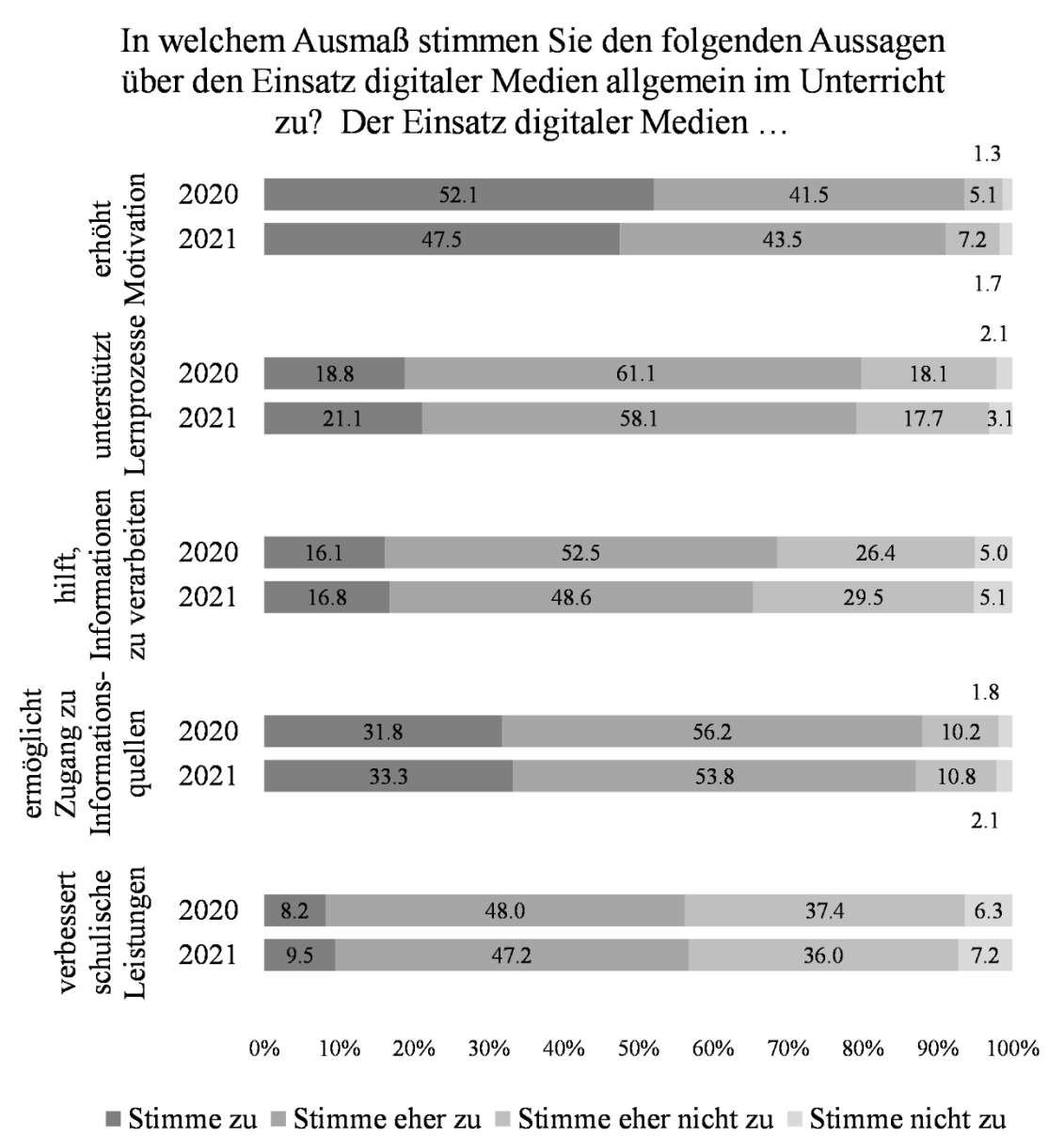
3.3 Einstellungen von Lehrkräften zum Einsatz digitaler Medien

Die Einstellungen von Lehrkräften sind, wie die bisherige Forschung zeigen konnte, ein Prädiktor dafür, dass und auf welche Weise digitale Medien im Lehr-Lernprozess integriert werden. Auf der Grundlage der vorliegenden Daten konnten die Angaben der Lehrpersonen zu den beiden Erhebungszeitpunkten verglichen werden, um der Frage nachzugehen, ob die Erfahrungen des verstärkt digital gestützten Unterrichts im Laufe der Pandemie mit Veränderungen in den Einstellungen der Lehrkräfte zum Einsatz digitaler Medien einhergehen.

Die Einstellungen der Lehrpersonen in diesem Kontext wurden auf der vierstufigen Antwortskala (1 = stimme zu; 4 = stimme nicht zu) erfasst. Abbildung 4 zeigt die Zustimmungsanteile der Lehrkräfte zu den Aussagen.

Abbildung 4

Einstellung der Lehrkräfte zum Einfluss digitaler Medien auf Schülerinnen und Schüler in Prozent



Notiz. Erhöht Motivation: „...erhöht die Motivation der Schülerinnen und Schüler.“; unterstützt Lernprozesse: „...unterstützt die Lernprozesse der Schülerinnen und Schüler.“; hilft, Informationen zu verarbeiten: „...hilft den Schülerinnen und Schülern, Informationen wirksamer zu vertiefen und zu verarbeiten.“; ermöglicht Zugang zu Informationsquellen: „...ermöglicht den Schülerinnen und Schülern den Zugang zu besseren

Informationsquellen.“; verbessert schulische Leistungen: „...verbessert die schulischen Leistungen der Schülerinnen und Schüler.“

Im Mittel stimmten die Teilnehmenden den Aussagen eher zu. Die Lehrkräfte waren somit eher der Ansicht, dass der Einsatz digitaler Medien die schulischen Leistungen der Schülerinnen und Schüler verbessert (2020: $M = 2.42$; $SD = 0.73$; 2021: $M = 2.41$; $SD = 0.76$), ihnen einen Zugang zu besseren Informationsquellen ermöglicht (2020: $M = 1.82$; $SD = 0.68$; 2021: $M = 1.82$; $SD = 0.70$), ihnen hilft, Informationen wirksamer zu vertiefen und zu verarbeiten (2020: $M = 2.20$; $SD = 0.77$; 2021: $M = 2.23$; $SD = 0.79$) und ihre Lernprozesse unterstützt (2020: $M = 2.03$; $SD = 0.68$; 2021: $M = 2.03$; $SD = 0.72$). Zwischen den Erhebungszeitpunkten waren keine statistisch signifikanten Unterschiede zwischen den Angaben der Lehrkräfte festzustellen. In Bezug auf die Motivation der Schülerinnen und Schüler sahen die meisten Lehrpersonen Potenziale und stimmten größtenteils zu, dass der Einsatz digitaler Medien die Motivation der Schülerinnen und Schüler erhöhte (2020: $M = 1.56$; $SD = 0.66$; 2021: $M = 1.63$; $SD = 0.70$). Im Vergleich der beiden Erhebungszeitpunkte lässt sich allerdings ein signifikanter Rückgang zum zweiten Erhebungszeitpunkt feststellen, wobei die diesbezügliche Einstellung der Lehrkräfte unter den erfassten Angaben weiterhin am positivsten ausfällt.

Zusammenfassend lassen sich durchschnittlich eher positive Einstellungen der Lehrkräfte zum Einsatz digitaler Medien im Unterricht aufzeigen. Im Laufe der Pandemie sind diese Einstellungen stabil geblieben, lediglich die sehr hohen motivationalen Potenziale wurden in geringerem Maße angegeben.

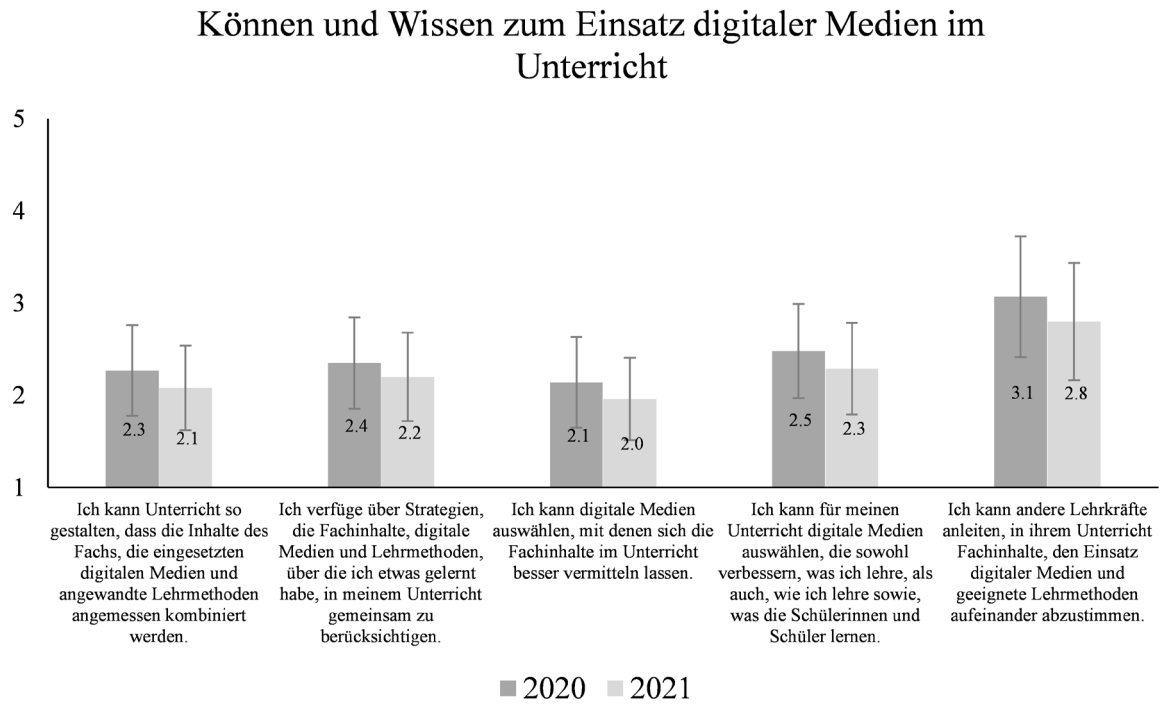
3.4 Kompetenzen der Lehrpersonen hinsichtlich des Einsatzes digitaler Medien in Lehr-Lernprozessen

Die Nutzung digitaler Medien für Lehr-Lernprozesse während der Corona-Pandemie erfordert eine Auseinandersetzung der Lehrkräfte mit digitalen Kommunikationswegen, digitalen Lernplattformen und digitalen Aufgabenformaten. Dabei stellte sich im Vergleich der beiden Erhebungszeitpunkte die Frage, wie Lehrkräfte ihre Kompetenzen hinsichtlich des Einsatzes digitaler Medien im Unterricht einschätzen und ob sich Unterschiede feststellen lassen.

Anknüpfend an das etablierte Instrument nach Schmidt et al. (2009) sind fünf Indikatoren zum technological pedagogical and content knowledge mit einem fünfstufigen Antwortformat erfasst worden (1 = stimme zu; 5 = stimme nicht zu).

Abbildung 5

Mittelwerte und Standardabweichungen zur Selbsteinschätzung der Lehrkräfte bezüglich der Kompetenz im Umgang mit digitalen Medien im Unterricht



Notiz. 1 = stimme zu; 5 = stimme nicht zu

Aus Abbildung 5 wird ersichtlich, dass die Lehrkräfte im Mittel den Aussagen eher zustimmten und damit ihre Kompetenzen zur Abstimmung der Lehrinhalte, der Lehrmethoden und des Einsatzes digitaler Medien insgesamt positiv einschätzten. Ihre Fähigkeit, andere Lehrkräfte dazu anleiten zu können, die Fachinhalte, den Einsatz digitaler Medien und geeignete Lehrmethoden aufeinander abzustimmen, schätzten die Lehrkräfte etwas zurückhaltender ein und gaben im Mittel an, teilweise dazu in der Lage zu sein. Für alle fünf erfassten Indikatoren zur Selbsteinschätzung der Kompetenz bezüglich des Einsatzes digitaler Medien lässt sich ein positiver Trend aufzeigen. Die Unterschiede zwischen den Erhebungszeitpunkten sind durchweg statistisch signifikant. Zu allen Aussagen war die Zustimmung der Lehrkräfte zum zweiten Erhebungszeitpunkt 2021 im Mittel höher.

Zusammenfassend kann hervorgehoben werden, dass die Erfahrungen des Medieneinsatzes während der Pandemie zu einer Wahrnehmung eines Kompetenzzuwachses bei den Lehrkräften beizutragen scheinen.

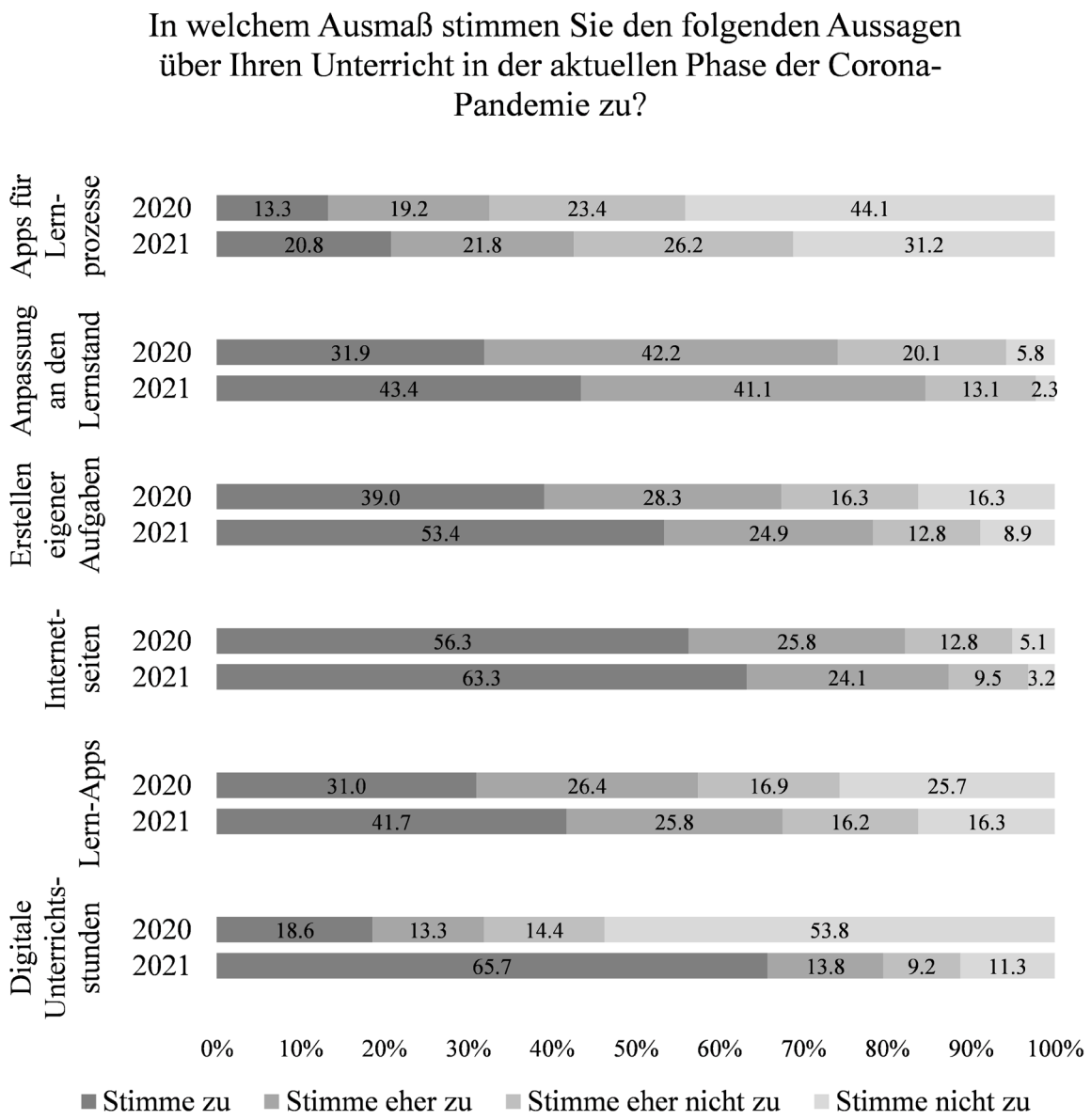
3.5 Lehren und Lernen mit digitalen Medien

Die bundesweite Befragung ermöglichte schließlich auch, die Nutzung digitaler Medien für die Organisation von Lehr-Lernprozessen zu betrachten. Die Pandemie stellte Lehrkräfte vor die Herausforderung, Unterricht an die gegebenen Umstände anzupassen, sodass das Lernen auch außerhalb des regulären Präsenzunterrichts stattfinden konnte und kann. Damit gingen beispielsweise Veränderungen von Aufgaben und Lernformaten einher, um ein selbstgesteuertes und eigenständiges Lernen zu Hause stärker zu unterstützen. Mit einem vierstufigen Antwortformat (1 = stimme zu; 4 = stimme nicht zu) wurde erfasst, wie Lehrkräfte digitale Medien verwendet haben, um Aufgaben und Lernprozesse während der Pandemie zu gestalten.

Die Durchführung von Unterrichtsstunden in digitaler Form (z.B. per Videokonferenz) stellte eine Option dar, auf die die Lehrkräfte in der zweiten Phase der Schulschließungen im Jahr 2021 deutlich intensiver zurückgriffen als noch ein Jahr zuvor. Während zu der Zeit der Schulschließungen im Frühjahr 2020 die Lehrkräfte im Mittel eher nicht zustimmten, digitale Unterrichtsstunden durchgeführt zu haben (M = 3.03; SD = 1.19), stimmte der Großteil der Lehrkräfte der Aussage ein Jahr später zu (M = 1.66; SD = 1.04). Die jeweiligen Anteile können Abbildung 6 entnommen werden.

Abbildung 6

Prozentuale Verteilung der Antworten zum digital gestützten Lehren und Lernen während der Schulschließungen.



Notiz. Apps für Lernprozesse: „Ich nutze Apps und digitale Anwendungen, mit denen die Schüler/innen ihren Lernprozess planen, dokumentieren und/oder reflektieren."; Anpassung an Lernstand: „Ich passe die Inhalte und Aufgaben auf den individuellen Lernstand meiner Schülerinnen und Schüler an."; Erstellen eigener Aufgaben: „Ich erstelle eigene digitale Aufgaben und modifiziere bestehende Ressourcen aus dem Internet, um sie an meine Bedarfe anzupassen."; Internetseiten: „Ich nutze verschiedene Internetseiten, um Unterrichtsmaterialen zu finden und auszuwählen."; Lern-Apps: „Ich habe Lern-Apps

genutzt, um die Lernprozesse der Schülerinnen und Schüler zu unterstützen."; digitale Unterrichtsstunden: „Ich führe Unterrichtsstunden digital (z.B. per Videokonferenz) durch."

Darüber hinaus hat die Nutzung von Lern-Apps zur Unterstützung der Lernprozesse der Schülerinnen und Schüler signifikant zugenommen (2020: $M = 2.37$; $SD = 1.17$; 2021: $M = 2.07$; $SD = 1.11$). Im Vergleich der Jahre ist festzustellen, dass die größten Veränderungen sich insbesondere in den Anteilen der Lehrkräfte zeigen, die der Aussage nicht zustimmten, sowie jener, die der Aussage zustimmten. Bereits zu Beginn der Pandemie nutzte ein Großteil der Lehrkräfte verschiedene Internetseiten, um Unterrichtsmaterialien zu finden, was der Umstellung des Unterrichts im Distanzformat zugutekommen kann. Nach einem Jahr ist der Anteil zum Zeitpunkt der bundesweiten Schulschließungen zu Beginn des Jahres 2021 im Mittel nochmals signifikant gestiegen (2020: $M = 1.66$; $SD = 0.89$; 2021: $M = 1.52$; $SD = 0.79$). Neben der Suche nach bereits bestehenden Materialien aus dem Internet kann es für den Lernstand und die gezielte Förderung der eigenen Lerngruppe auch erforderlich sein, eigene digitale Aufgaben zu erstellen oder bestehende Materialien zu modifizieren. Auch zu diesem Vorgehen zeigte der Vergleich der beiden Erhebungszeitpunkte eine intensivere individuelle Aufgabengestaltung durch die Lehrkräfte (2020: $M = 2.10$; $SD = 1.09$; 2021: $M = 1.77$; $SD = 0.98$). Zudem stimmten die Lehrkräfte im Jahr 2020 im Mittel eher zu, die Inhalte und Aufgaben auf den individuellen Lernstand der Schülerinnen und Schüler anzupassen (2020: $M = 2.00$; $SD = 0.87$). Diesbezüglich lässt sich ein Jahr später ebenfalls ein signifikanter Unterschied mit einer durchschnittlich höheren Zustimmung feststellen (2021: $M = 1.74$; $SD = 0.77$). Schließlich wurde erfasst, ob die Lehrkräfte auch Apps und digitale Anwendungen herangezogen haben, mit denen das selbstgesteuerte Lernen der Schülerinnen und Schüler begleitet wird, sodass die Lernenden selbst ihre Lernprozesse planen, dokumentieren und reflektieren. Im Mittel stimmten die Lehrkräfte dieser Aussage eher nicht zu, wobei die Anteile der zustimmenden oder eher zustimmenden Lehrkräfte zum zweiten Erhebungszeitpunkt höher ausfielen (2020: $M = 2.98$; $SD = 1.08$; 2021: $M = 2.68$; $SD = 1.12$). Insgesamt kann festgestellt werden, dass die Nutzung digitaler Medien für die Organisation der Lehr-Lernprozesse im Laufe der Pandemie zugenommen hat. Für alle berichteten Indikatoren konnte ein signifikanter Unterschied zwischen den Erhebungszeitpunkten festgestellt werden, der jeweils auf eine höhere Zustimmung digital gestützter Vorgehensweisen zum Zeitpunkt der Erhebung 2021 hinweist.

4. Diskussion und Ausblick

Die vorliegende Studie untersuchte anhand zweier Online-Lehrkräftebefragungen, wie mit der außergewöhnlichen Situation der pandemiebedingten Schulschließungen ab März 2020 sowie zu Beginn des Jahres 2021 umgegangen wurde und welche Rolle digitale Medien dabei einnahmen. Besonders interessant ist dabei der Vergleich der beiden Phasen der Schulschließungen, der Hinweise auf Implementationen und Trends hinsichtlich der Nutzung digitaler Medien für die Aufrechterhaltung der Kommunikation und des Unterrichts lieferte und Veränderung in den Einstellungen und Selbsteinschätzung der Kompetenzen von Lehrkräften bezüglich des Medieneinsatzes aufzeigte. Zu den zentralen Herausforderungen gehörte dabei, dass plötzlich verstärkt digitale Medien herangezogen wurden, die zuvor im deutschen Bildungssystem nicht umfassend etabliert waren (Eickelmann et al., 2019; Lorenz et al., 2017).

Die empirischen Befunde zeigten, dass sich das Lernen und Lehren mit digitalen Medien zwischen der ersten bundesweiten Schulschließung 2020 und der zweiten zu Beginn des Jahres 2021 in vielerlei Hinsicht gewandelt hat. Hinsichtlich der technischen Grundvoraussetzungen ist vor allem eine Verbesserung der schülerseitigen Ausstattung zu verzeichnen. Ebenfalls gibt es für Schülerinnen und Schüler vermehrt Möglichkeiten zum Geräteausleih von der Schule. Diese positive Entwicklung ist insbesondere dahingehend relevant, die Bildungsbeteiligung aller Schülerinnen und Schüler in diesen herausfordernden Zeiten zu sichern. Wechselunterricht und hybride Formen werden ebenfalls von diesen Maßnahmen profitieren. Mit Blick auf die Zufriedenheit der Lehrpersonen mit der technischen Ausstattung und dem Auftreten von Technikproblemen wurde jedoch weiterhin Optimierungspotenzial deutlich. Mit einer bedarfsgerechten Ausstattung, die den pädagogischen Anforderungen gerecht wird, kann die Qualität des Unterrichts unterstützt werden.

Hinsichtlich der Kommunikation zeigte sich, dass Lehrkräfte weniger E-Mails oder das Telefon nutzten und stattdessen vermehrt Lernplattformen und virtuelle Treffen einsetzten, um Schülerinnen und Schüler zu erreichen. Dieser Trend weg von individuellen Kontaktmedien und hin zu virtuellem Unterricht hin wurde bereits nach den Osterferien 2020 vermutet (BiSE, 2020). Schon vor der Krise nutzten Lehrkräfte individualisierte Kontaktmöglichkeiten wie E-Mails oder das Telefon zur Kommunikation mit Schülerinnen und Schülern (vgl. ebd.). Ein weiterer Einsatz dieser Kommunikationsformen zu Beginn der Schulschließungen lässt sich daher vermutlich darauf zurückführen, dass Lehrkräfte ihnen

bereits vertraute Kommunikationsformen zunächst weiter nutzen wollten. Dem gegenüber standen mit Lernplattformen und virtuellem Unterricht Werkzeuge, mit deren Nutzung Lehrkräfte vor den Schulschließungen weniger vertraut waren oder die nicht an allen Schulen zur Verfügung standen. Durch die Schulschließungen stiegen Nachfrage und Angebot von digitalen Lern- und Lehrmedien, sodass Gruppenkommunikationsmedien prävalenter eingesetzt werden konnten. Während der zweiten bundesweiten Schulschließung wurde Unterricht von einem Großteil der Lehrkräfte virtuell abgehalten und auch Lernplattformen wurden von der breiten Mehrheit genutzt. Neben der mangelnden Vertrautheit boten auch Unklarheiten über datenschutzrechtliche Aspekte eine Hürde für einen breiteren Einsatz von virtuellen Konferenzprogrammen und Lernplattformen, welche im Laufe der Corona-Pandemie zunehmend flächendeckend zur Verfügung gestellt wurden.

Des Weiteren wurden die Einstellungen der Lehrkräfte zu digitalen Medien im Unterricht betrachtet, die sich zwischen den Erhebungen kaum unterschieden. Dabei ist zu bedenken, dass ein Großteil der Lehrkräfte digitale Medien zur Informationssuche und Vertiefung von Lerninhalten von Schülerinnen und Schülern bereits grundsätzlich positiv bewertete. Obgleich die Einstellung zu digitalen Medien ein wichtiger Prädiktor für deren Einsatz im Unterricht darstellt (Celik & Yesilyurt, 2013; Ertmer, 2005; Scherer et al., 2020; Siyam, 2019), ließ sich zwischen den Erhebungen zwar ein Anstieg in der Nutzung, aber keine Veränderung in den Einstellungen finden. Dies lässt sich hauptsächlich mit den außergewöhnlichen Umständen, die zu dem Einsatz der digitalen Medien führen, erklären. So sind Lehrkräfte im Distanzunterricht angehalten, von digitalen Medien auch in Lehr-Lernsituationen Gebrauch zu machen, die im Präsenzunterricht ohne digitale Medien auskommen. Zudem lassen sich Probleme mit der Technik oder dem Internet, die Lehrkräfte nach wie vor im Unterricht behindern, im Distanzunterricht nicht umgehen. Sie müssen im Unterrichtsverlauf und in der aktuellen Situation zwingend behoben werden. Darin sind möglicherweise Gründe dafür zu sehen, dass keine Unterschiede in den Einstellungen der Lehrkräfte zum Medieneinsatz zu verzeichnen sind.

Hervorgehoben werden kann in diesem Kontext, dass Lehrkräfte über den Verlauf des Distanzunterrichts eine eher geringere Motivation der Schülerinnen und Schüler durch den Einsatz digitaler Medien empfanden. Ein Grund dafür könnte sein, dass digitale Medien im Regelunterricht seltener zum Einsatz kamen und dann eine besondere Unterrichtsstunde signalisierten, durch die die Schülerinnen und Schüler stärker interessiert und motiviert

wurden. Durch den Distanz- und Wechselunterricht sind digitale Medien hingegen eher zur Gewohnheit geworden und der motivierende Effekt könnte daher abgemildert sein.
Die Erfahrungen der Lehrkräfte mit dem Einsatz digitaler Medien in Lehr- und Lernsituationen tangierten ihre medienbezogenen Kompetenzen, indem eine intensivere Nutzung und Auseinandersetzung mit digitalen Medien gefordert waren, deren Potenziale für synchrone und asynchrone Lernsituationen ausgelotet werden und Lernräume zeitlich und räumlich geöffnet sowie individualisiertes Lernen stärker beachtet werden mussten. Die Befunde geben Hinweise auf eine höhere Selbstwirksamkeit der Lehrkräfte, die sich nach einem Jahr des Unterrichts unter Pandemie-Bedingungen im Mittel vermehrt Kompetenzen im Unterrichten mit digitalen Medien zuschrieben. Diese Erfahrungen und damit einhergehende Kompetenzen müssen nun systematisiert und damit für eine nachhaltige und lernförderliche Nutzung digitaler Medien und zukunftsfähige Gestaltung des Unterrichts genutzt werden.

Aktives Nutzen von digitalen Medien im Unterricht ist eine Strategie, um die technologischen, pädagogischen und (fach-)inhaltlichen Wissensdomänen zu erhöhen (Tondeur et al., 2019). Damit im Einklang steht auch der Befund einer vermehrten Nutzung digitaler Medien für die Unterrichtsorganisation. Neben der vermehrten Nutzung von Konferenzprogrammen stieg nicht nur der Anteil an Lehrkräften, der virtuellen Unterricht durchführte, zwischen den Zeitpunkten der Schulschließungen an. Lehrkräfte nutzten Anfang 2021 zudem verstärkt das Internet, um nach Aufgaben für ihre Klassen zu suchen. Diese Aufgaben wurden intensiver modifiziert und an individuelle Bedürfnisse angepasst. Auch Lern-Apps wurden aktiver eingesetzt. Aufgrund ihrer Erfahrungen im und mit dem Distanzunterricht ist es demnach nicht verwunderlich, dass sich die Lehrkräfte im Umgang mit digitalen Medien als kompetenter ansahen, wenn sie Medien verstärkt für den digitalen Unterricht nutzten.

Zusammenfassend zeichnen die Befunde eine Intensivierung der Digitalisierung im Distanzunterricht nach einem Jahr der Pandemie ab, die jedoch weiterhin von Hemmnissen geprägt ist. Zudem wird ersichtlich, dass zahlreiche Erfahrungen und Nutzungsweisen mit digitalen Medien gesammelt wurden. Eine mögliche Implikation lässt sich dahingehend ableiten, dass gute Praxisbeispiele systematisiert und empirisch hinsichtlich der Förderung der Schülerleistungen evaluiert werden sollten. Auf diesem Wege können wertvolle Erkenntnisse für den digital gestützten Unterricht gewonnen werden, die auch für die Zeit nach der Pandemie gewinnbringend sein können. Für die Kompetenzentwicklung der Lehrkräfte ist

zudem relevant zu untersuchen, worauf die durchschnittlich höhere Einschätzung der Kompetenzen im Einsatz digitaler Medien für den Unterricht zurückzuführen ist. Eng damit verbunden ist auch die Frage, ob die Erstellung, Suche und Modifikation von Materialien für den digital gestützten Unterricht zunehmend professionalisierter ablaufen – beispielsweise die Einschätzung der Qualität der Materialien kriteriengeleitet vorgenommen wird – oder lediglich schnellere Routinen vorzufinden sind.

Eine Limitation der vorliegenden Studie ist, dass die vorliegenden Daten nicht im Längsschnitt bei den gleichen Teilnehmenden erhoben werden konnten, sodass Entwicklungen und Effekte analysiert werden können. Stärken der Untersuchung liegen in der hohen Zahl der Teilnehmenden aus allen Bundesländern, die mit beiden Befragungen erreicht werden konnten. Somit kann ein umfassender Gesamteindruck der Unterrichtssituation während der Pandemie und der Rolle digitaler Medien gewonnen werden.

Überblick über bisherige Publikationen zur Studie und Website

Website der Studie:
http://www.ifs.tu-dortmund.de/cms/de/Forschung/Gesamtliste-Laufende-Projekte/Corona-U-2021.html

Bisher sind im Rahmen der Studie folgende Publikationen veröffentlicht worden, die auf Analysen des ersten erhobenen Datensatzes zur Zeit der Schulschließungen im Frühjahr 2020 basieren:

McElvany, N., Lepper, C., Lorenz, R. & Brüggemann, T. (2021). Unterricht während der Corona-Pandemie. In D. Dohmen & K. Hurrelmann (Hrsg.), *Generation Corona? Wie Jugendliche durch die Pandemie benachteiligt werden* (S. 64–79). Weinheim: Beltz Juventa.

Auf Basis der bundesweiten Lehrkräftebefragung zum Unterricht während der Corona-Pandemie untersuchte der Beitrag die Kommunikation zwischen Lehrkräften und Lernenden während den bundesweiten Schulschließungen im Frühjahr 2020. Zudem standen Aspekte der technischen Rahmenbedingungen, der Unterrichtsgestaltung, der Zusammenarbeit mit Eltern sowie Einschätzungen zum Lernerfolg der Schülerinnen und Schüler im Fokus. Hervorzuheben ist der Vergleich der Angaben von Lehrkräften

verschiedener Schulformen, sodass schulformspezifische Betrachtungen ermöglicht wurden. Die Erfahrungen der befragten Lehrkräfte – gerade auch zu unterschiedlichen Gegebenheiten der Schulformen in Bezug auf eine Vielzahl der untersuchten Aspekte – können als zentraler Informationsbaustein für Bildungsadministration, Politik und Öffentlichkeit erachtet werden.

Lorenz, R., Lepper, C., Brüggemann, T., & McElvany, N. (2020). U*nterricht während der Corona-Pandemie. Lehrkräftebefragung. Ergebnisse, Teil I: „Der Unterricht“*. Dortmund: Institut für Schulentwicklungsforschung (IFS).

Im Fokus des Beitrags steht der Unterricht während der Corona-Pandemie zum Zeitpunkt der Schulschließungen im Frühjahr 2020 in Deutschland. Die Lehrkräftebefragung zeigte, dass die technische Ausstattung sowie die Medienkompetenz der Lernenden nach Einschätzung der Lehrkräfte eher nicht ausreichend waren, um den Unterricht digital gestützt fortzuführen, insbesondere bei Grundschulkindern. Außerdem mangelte es bei möglichen Bedarfen an Ausleihmöglichkeiten digitaler Medien von der Schule. Die Kommunikation und Übermittlung von Aufgaben erfolgten eher uneinheitlich auf verschiedenen Kommunikationswegen. Eine große Mehrheit der Lehrkräfte stellte den Lernenden digital zu bearbeitende Aufgaben und Materialien zur Verfügung und passte nach eigener Auskunft Aufgaben an individuelle Lernstände der Schülerinnen und Schüler an. Die Zusammenarbeit mit den Eltern wurde von Lehrkräften aller Schulformen eher positiv eingeschätzt. Die Lehrkräfte berichteten in großer Mehrheit, dass die Kinder in vielen Fächern weniger als normalerweise in der Schule gelernt haben und dass die sozial bedingten Ungleichheiten im Bildungsbereich verstärkt wurden.

Lorenz, R., Brüggemann, T., & McElvany, N. (2020). *Unterricht während der Corona-Pandemie. Zweiter Ergebnisbericht der bundesweiten Lehrkräftebefragung. Ergebnisse, Teil II: „Wohlbefinden der Lehrkräfte“*. Dortmund: Institut für Schulentwicklungsforschung (IFS).

Das Wohlbefinden der Lehrkräfte während der ersten Phase der Schulschließungen stellte im Beitrag ein besonderes Interesse dar. Es wurde deutlich, dass für 42.0 Prozent der Lehrkräfte gleichzeitig zu den beruflichen Herausforderungen während der Corona-Pandemie die Kinderbetreuung sicherzustellen war, was insbesondere angesichts fortdauernder

Ungewissheiten hinsichtlich des Regelbetriebs der Schulen von Relevanz war. Die häusliche Situation der Lehrkräfte in den Wochen der coronabedingten Schulschließungen wurde bezogen auf die Arbeitsbedingungen insgesamt als mittelmäßig eingeschätzt, wobei jedoch deutliche Unterschiede zwischen den Lehrkräften ersichtlich waren. Die Gefühlslage der Lehrkräfte im Alltag war zu der Situation rund um die Corona-Pandemie insgesamt eher ausgewogen. Positive und negative Affekte wurden etwa gleichermaßen empfunden, wobei insbesondere die Unterstützung durch das Kollegium, die Schulleitung sowie die Eltern relevant für das Wohlbefinden der Lehrkräfte war. Die Berufszufriedenheit der Lehrkräfte in der Zeit während der Corona-Pandemie war eher hoch, eine Wertschätzung ihrer Arbeit haben sie im Vergleich dazu in geringerem Maße empfunden. Die empfundene Belastung und der Stress wurden von den Lehrkräften im Durchschnitt auf einem mittleren Niveau eingestuft. Insbesondere weibliche Lehrkräfte und Lehrkräfte mit mangelnder Unterstützung seitens der Schulleitung und der Eltern hatten ein höheres Belastungs- und Stressempfinden.

6. Literatur

Anger, S., Bernhard, S., Dietrich, H., Lerche, A., Patzina, A., Sandner, M., et al. (2020). *Schulschließungen wegen Corona: Regelmäßiger Kontakt zur Schule kann die schulischen Aktivitäten der Jugendlichen erhöhen*. IAB-Forum. Verfügbar unter: https://www.iab-forum.de/schulschliessungen-wegen-corona-regelmassiger-kontakt-zur-schule-kann-die-schulischen-aktivitaten-der-jugendlichen-erhohen/ [Zugriff am 29.04.2021]

BiSE. (2020). *Lehrerbefragung zur Schulschließung*. Universität Konstanz. Verfügbar unter: https://www.uni-konstanz.de/typo3temp/secure_downloads/64520/0/f9e3cdeced59dfd07f480d947048a2c99eca5a38/Kurzbericht_Lehrerbefragung_Schulschlie%C3%9Fung_final.pdf [Zugriff am 29.04.2021].

BMBF [Bundesministerium für Bildung und Forschung] (Hrsg.) (2016). *Bildungsoffensive für die digitale Wissensgesellschaft – Strategie des Bundesministeriums für Bildung und Forschung*. Berlin: BMBF.

BMBF [Bundesministerium für Bildung und Forschung] (2019). *Verwaltungsvereinbarung DigitalPakt Schule 2019 bis 2024*. Verfügbar unter: https://www.digitalpaktschule.de/files/VV_DigitalPaktSchule_Web.pdf [Zugriff am 23.04.2021]

Celik, V., & Yesilyurt, E. (2013). Attitudes to technology, perceived computer self-efficacy and computer anxiety as predictors of computer supported education. *Computers & Education*, *60*(1), 148–158. https://doi.org/10.1016/j.compedu.2012.06.008

Davis, N., Eickelmann, B. & Zaka, P. (2013). Restructuring of educational systems in the digital age from a co-evolutionary perspective. *Journal of Computer-Assisted Learning 29*, 438–450.

Drent, M. & Meelissen, M. (2008). Which factors obstruct or stimulate teacher educators to use ICT innovatively? *Computers & Education*, *51*(1), 187–199. https://doi.org/10.1016/j.compedu.2007.05.001

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., et al. (Hrsg.) (2019*). ICILS 2018 #Deutschland – Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking*. Münster: Waxmann.

Eickelmann, B. & Vennemann, M. (2017). Teachers‘ attitudes and beliefs regarding ICT in teaching and learning in European countries. *European Educational Research Journal*, *16*(6), 733–761. https://doi.org/10.1177/1474904117725899

Ertmer, P.A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational Technology Research and Development*, *53*(4), 25–39. https://doi.org/10.1007/BF02504683

Fickermann, D. & Edelstein, B. (2020). „Langsam vermisse ich die Schule ...“ Schule während und nach der Corona-Pandemie. *DDS – Die Deutsche Schule, Beiheft 16*, 9–36.

Huber, S.G., Günther, P.S., Schneider, N., Helm, C., Schwander, M., Schneider, J.A., et al. (2020). *COVID-19 und aktuelle Herausforderungen in Schule und Bildung. Erste Befunde des Schul-Barometers in Deutschland, Österreich und der Schweiz.* Münster: Waxmann.

KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland]. (2016). *Bildung in der digitalen Welt. Strategie der Kultusministerkonferenz*. Verfügbar unter: https://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf [Zugriff am 23.04.2021]

Knezek, G. & Christensen, R. (2018). The Evolving Role of Attitudes and Competencies in Information and Communication Technology in Education. In J. Voogt, G. Knezek, R. Christensen, & K. Lai (Hrsg.), *Second Handbook of Information Technology in Primary and Secondary Education* (pp. 239–253). Springer. https://doi.org/10.1007/978-3-319-71054-9_16

Köller, O., Fleckenstein, J., Guill, K. & Meyer, J. (2020). Pädagogische und didaktische Anforderungen an die häusliche Aufgabenbearbeitung. *DDS – Die Deutsche Schule, Beiheft 16*, 163–176.

Lorenz, R. & Endberg, M. (2017). IT-Ausstattung der Schulen der Sekundarstufe I im Bundesländervergleich und im Trend von 2015 bis 2017. In R. Lorenz, W. Bos, M. Endberg, B. Eickelmann, S. Grafe & J. Vahrenhold (Hrsg.), *Schule digital – der Länderindikator 2017. Schulische Medienbildung in der Sekundarstufe I mit besonderem Fokus auf MINT-Fächer im Bundesländervergleich und Trends von 2015 bis 2017* (S. 49–83). Münster: Waxmann.

Lorenz, R., Endberg, M. & Bos, W. (2019). Predictors of fostering students' computer and information literacy – analysis based on a representative sample of secondary school teachers in Germany. *Education and Information Technologies*, *24*(1), 911–928. https://doi.org/10.1007/s10639-018-9809-0

Lorenz, R., Lepper, C., Brüggemann, T. & McElvany, N. (2020). U*nterricht während der Corona-Pandemie. Lehrkräftebefragung. Ergebnisse, Teil I: „Der Unterricht"*. Dortmund: Institut für Schulentwicklungsforschung (IFS). Verfügbar unter: http://www.ifs.tu-dortmund.de/cms/de/Home/Pressematerialien/Pressematerialien/UCP_Kurzbericht_final.pdf [Zugriff am 27.04.2021]

McElvany, N. (2018). Digitale Medien in den Schulen: Perspektive der Bildungsforschung. In N. McElvany, F. Schwabe, W. Bos & H. G. Holtappels (Hrsg.), *Digitalisierung in der schulischen Bildung: Chancen und Herausforderungen* (S. 99–106). Münster: Waxmann.

Mishra, P. & Koehler, M.J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Sander, A., Schäfer, L., & van Ophuysen, S. (2020). *Erste Ergebnisse aus dem Projekt „Familiäre Lernbegleitung in Zeiten von Schulschließungen aufgrund der Corona-Pandemie (FamiLeb)"*. Westfälische Wilhelms-Universität Münster: Institut für Erziehungswissenschaft.

Scherer, R., Siddiq, F., & Tondeur, J. (2020). All the same or different? Revisiting measures of teachers' technology acceptance. *Computers & Education*, *143*, 103656. https://doi.org/10.1016/j.compedu.2019.103656

Schmidt, D.A., Baran, E., Thompson, A.D., Mishra, P., Koehler, M.J. & Shin, T.S. (2009). Technological Pedagogical Content Knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 24*(2), 123–149.

Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre- and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, *115*, 106552. https://doi.org/10.1016/j.chb.2020.106552

Siyam, N. (2019). Factors impacting special education teachers' acceptance and actual use of technology. *Education and Information Technologies*, *24*(3), 2035–2057. https://doi.org/10.1007/s10639-018-09859-y

Steinmayr, R., Lazarides, R., Weidinger, A.F. & Christiansen, H. (2021). Teaching and learning during the first COVID-19 school lockdown: Realization and associations with parent-perceived students' academic outcomes. *Zeitschrift für Pädagogische Psychologie* (2021), 1–22. https://doi.org/10.1024/1010-0652/a000306

Tondeur, J., Scherer, R., Siddiq, F. & Baran, E. (2019). Strategies to prepare pre-service teachers for Technological Pedagogical Content Knowledge (TPACK): A mixed-method study. *Educational Technology Research & Development* 68, 319–343. https://doi.org/10.1007/s11423-019-09692-1

Tondeur, J., van Braak, J., Ertmer, P.A. & Ottenbreit-Leftwich, A. (2016). Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence. *Educational Technology Research and Development*, *65*(3), 555–575. https://doi.org/10.1007/s11423-016-9481-2https://doi.org/10.1007/s11423-016-9481-2

van Ackeren, I., Aufenanger, S., Eickelmann, B., Friedrich, F., Kammerl, R., Knopf, J., et al. (2019). Digitalisierung in der Lehrerbildung. Herausforderungen, Entwicklungsfelder und Förderung von Gesamtkonzepten. *Die Deutsche Schule 111*, 103–119.

vbw – Vereinigung der Bayerischen Wirtschaft e. V. (Bos, W., Daniel, H.-D., Hannover, B., Köller, O., Lenzen, D., McElvany, N., Roßbach, H.-G., Seidel, T., Tippelt, R., Wößmann, L., Mitglieder des Aktionsrats Bildung) (2018). *Digitale Souveränität und Bildung*. Gutachten. Münster: Waxmann.

Voss, T. & Wittwer, J. (2020). Unterricht in Zeiten von Corona: Ein Blick auf die Herausforderungen aus der Sicht von Unterrichts- und Instruktionsforschung. *Unterrichtswissenschaft, 48*, 601–627.

Wößmann, L., Freundl, V., Grewenig, E., Lergetporer, P., Werner, K. & Zierow, L. (2020). Bildung in der Coronakrise: Wie haben die Schulkinder die Zeit der Schulschließungen verbracht, und welche Bildungsmaßnahmen befürworten die Deutschen? *ifo Schnelldienst, 73*(9), 25–39.

3.2 Beitrag II: Effects of Mode and Medium in Reading comprehension Tests on Cognitive Load

Brüggemann, T., Ludewig, U., Lorenz, R., & McElvany, N. (2023). Effects of Mode and Medium in Reading comprehension Tests on Cognitive Load. *Computers & Education*, 192. https://doi.org/10.1016/j.compedu.2022.104649

Dieses Kapitel enthält keine Kopie des abgedruckten Artikels und weicht von der publizierten Version leicht ab.

Zusammenfassung

Digitale Medien erlauben den Einsatz von computeradaptive Tests (CAT), die sich an die Fähigkeiten der Testpersonen anpassen. Computeradaptive- und computerbasierte Tests (CBT) unterscheiden sich von papierbasierten Tests (PPT) in ihrer Darstellung. Im Rahmen der *cognitive load theory* können sich Unterschiede in der Darstellungsart (z.B. Papier und Bildschirm) auf die kognitive Belastung auswirken, die bei einem Lesekompetenztest empfunden wird. Aufgrund der Anpassung der Aufgabenschwierigkeit in CATs an die Testperson ist es im Rahmen der *working memory resource depletion* naheliegend, dass die kognitive Belastung in CATs stärker zunimmt. In dieser Studie wurde untersucht, wie sich die Testformate PPT, CBT und CAT auf die kognitive Belastung von 212 Schüler*innen der vierten Klassen auswirkt. Schüler*innen wurden zufällig einer der drei Testformaten zugewiesen und in der Mitte und am Ende eines Lesekompetenztests nach ihrer kognitiven Belastung gefragt. Ein linear mixed-effects model fand keine Unterschiede zwischen den Testformaten hinsichtlich der kognitiven Belastung, jedoch gab es Hinweise auf einen erhöhten Anstieg der kognitiven Belastung im CAT gegenüber dem PPT und CBT.

Abstract

Digital media are becoming increasingly prevalent in the assessment of student achievement. Assessments that use digital media rather than pencil-and-paper (PPT) present the test materials on a screen (computer-based test: CBT) and can apply adaptive testing procedures (computer adaptive test: CAT). Based on cognitive load theory, presentation differences between the screen and paper might impact the cognitive load experienced during a reading comprehension test. In addition, working memory resource depletion might be indicative of an accumulation of cognitive load over the course of a test. Due to their active alignment of item difficulty and person ability, cognitive load might increase more in CATs than in PPTs or CBTs. Research has mostly compared test scores resulting from these different test formats, while systematic differences in the testing experience, such as experienced cognitive load, have received less scholarly attention. This study investigated how the three test formats of PPT, CBT, CAT affect examinees' level of experienced cognitive load in the middle and at the end of a reading comprehension test. In a between- and within-subject design, 212 German fourth graders (age: $M = 9.44$, $SD = 0.59$) were randomly assigned to a standardized reading comprehension test administered in one of the three test formats (between-subject: CBT, CAT, or PPT). Linear mixed-effects models revealed no significant mean differences in experienced cognitive load between the three test conditions, but a higher rate of increase in cognitive load in the adaptive test, although the effects were relatively small. Implications for future reading comprehension assessments using digital media are discussed.

1. Introduction

Monitoring students' competences is vital in order to inform teachers, administrators, and policymakers about students' learning progress and achievement heterogeneity. For educational researchers, it is a matter of accountability to use state of the art designs to make assessments as reliable, valid and efficient as possible. Digital media have recently been described as a "third space" for learning, located between learning at home and at school (McDougall & Potter, 2019; Potter & McDougall, 2017). This view has gained new relevance in light of the COVID-19 pandemic, which forced schools across the globe into distance learning and thus further increased reliance on computers for schooling (Helm et al., 2021). Large-scale assessments such as PIRLS and PISA have started to incorporate computers into their test administration (Hußmann et al., 2017; Yamamoto et al., 2019). Administering tests on the computer comes with the option to use testing modes that adapt to each examinee's ability (Davey, 2011; Frey et al., 2017). Reading comprehension assessments in elementary school are of particular interest, because reading comprehension affects students' educational futures and their ability to participate in society and life in general (OECD, 2019; Wigfield et al., 2016). As such, it is one of the most important skills that children are taught in elementary school. There, children move from learning individual letters to decoding the meaning of words, on to understanding the content of a sentence, paragraph, or text (Becker, McElvany, & Kortenbruck, 2010). Between grades two and three, children tend to begin reading fluently (see Chall, 1983). Still, reading ability in fourth grade students is considerably heterogeneous, with only 34% of US students participating in the National Assessment of Educational Progress (NAEP) reaching reading proficiency (as defined by the NAEP), and 35% not reaching the basic reading level (National Center for Education Statistics, 2019). Computer-based tests (CBTs) and computer-adaptive tests (CATs) can potentially improve the assessment of reading comprehension over paper-and-pencil tests (PPTs). CBTs have higher technical demands, but offer more control over the test situation (i.e., exposure control or time limits) and provide additional quality control data (e.g., rapid guessing detection) and process data (e.g., log files) compared to PPTs. In addition, adaptive item selection can make a test design much more efficient (Davey, 2011). However, for beginning readers, the effects of reading on screen versus reading on paper and the potential effects of adaptive item selection that adjusts to individual ability are not well-researched. Existing research has focused on test score equivalence between different formats of test administration regarding test scores. Research into the *test experience* has been limited. Inspired by basic research on instructional

design and cognitive psychology, it has been suggested that taking a CBT, CAT, or PPT can lead to different test experiences (Colwell, 2013; Ortner et al., 2014). An important aspect of the test experience is an examinee's cognitive load and its relation to working memory. Reading comprehension is closely related to working memory (Daneman & Carpenter, 1980), as comprehension relies on remembering previous words in a sentence or paragraph. This holds true for elementary school children (Seigneuric et al., 2000). Cognitive load is the strain that carrying out a task, such as learning or reading, puts on a person's working memory (Sweller et al., 2019). Cognitive load theory (CLT) is often applied in instructional design in an attempt to optimize the cognitive load of instructions in order to improve learning (Sweller, 1988; Sweller et al., 2019). CLT attempts to explain the transfer of information between working memory and long-term memory. Cognitive load theory distinguishes between three different types of cognitive load. Intrinsic cognitive load is inherent and specific to the complexity of a task (Sweller et al., 2019). The more complex a task, i.e., the more interconnected and interactive elements it contains, the more intrinsic cognitive load a person experiences when working on the task. Extraneous load depends on the presentation of the learning material (Sweller et al., 2019). For instance, difficult-to-read fonts can increase extraneous cognitive load (Ball et al., 2018). Lastly, germane cognitive load is a positive type of load in which intrinsic load is dealt with by "redistributing working memory resources from extraneous activities to activities directly relevant to learning" (Sweller et al., 2019). Cognitive load is conceptually (Leppink et al., 2014) and empirically (Mayes et al., 2001; van de Weijer-Bergsma & van der Ven, 2021) negatively associated with test performance.

Additionally, cognitive load may accumulate over the course of a test administration situation. Chen et al. (2018) incorporated *working memory resource depletion* into the cognitive load framework (see also Leahy & Sweller, 2019). They showed that fourth grade students perform better on a working memory test and on a subsequent mathematics test when spacing out learning sessions over three days compared to students whose learning sessions took place over a single day. They conclude that working memory is a limited resource that depletes when exerting cognitive effort for a given task. Hence, mental effort on one task diminishes performance on successive tasks that are similar to the original task as working memory depletes.

With test formats differing with regard to the medium of administration – paper or screen – as well as with regard to the test mode – adaptive or fixed – different test administration situations may differentially affect experienced cognitive load in a test

situations. This study investigates the effects the different test formats PPT, CBT, and CAT have on the cognitive load of fourth grade students over the course of a reading comprehension test.

2. Test Formats and Cognitive Load

2.1 Test Medium: Paper-and-Pencil Tests (PPT) and Computer-Based Tests (CBT)

In many educational contexts, PPTs are still the prevalent medium of test administration (Martin & Lazendic, 2018). PPTs have low technical demands, because paper test booklets are easy to create, edit and administer. They require little technical knowledge of both administrators and examinees. Still, computer-based tests (CBT) have become more popular in recent years, as scoring can be automated and response times and log-files can be assessed (Alruwais et al., 2018).

Many studies have compared test outcomes in terms of students' reading comprehension scores when reading on screen versus reading on paper. While some studies found no differences (Porion et al., 2016), multiple recent meta-analyses found a small but significant "screen inferiority" effect (Clinton, 2019; Delgado et al., 2018; Kong et al., 2018), indicating that reading comprehension suffers when reading on screen. This effect is more or less pronounced depending on factors such as text type (Clinton, 2019), device, or reading time (Delgado et al., 2018). In a moderator analysis, Delgado et al. (2018) also concluded that the screen inferiority effect was smaller and only marginally significant when scrolling was not necessary. Screen inferiority also occurs among participants aged 15-16 (Mangen et al., 2013) and aged 10 (Støle et al., 2020), but is not dependent on students' computer familiarity or medium preferences (Halamish & Elbaz, 2020). Nevertheless, differences between reading on screen and on paper may depend on the aforementioned moderators, as some studies found evidence for the equivalence of tests on screen and paper when eliminating certain aspects of screen reading like the need for scrolling (Porion et al., 2016), which can affect readers' engagement (Mangen & Kuiken, 2014) with the text and its mental representation (Piolat et al., 1997). Furthermore, decisions in designing the user interface can affect test performance. For instance, a study by Buerger et al. (2019) found that using drop-down menus to answer an item instead of a multiple choice format increased item difficulty, as well as placing texts on a different page than the items, since examinees needed to switch back and forth between text and item. Similarly, in a study with fourth-grade students by Dawidowsky et al. (2021),

reading comprehension test scores were higher in a tablet-based CBT when the user interface was optimized for tablet use than in a PPT, while no difference was found between the PPT and the CBT with a non-optimized user interface.

2.2 Cognitive Load when Reading on Paper and on Screen

Differences in how humans perceive text on screen compared to text on paper may affect readers' experienced cognitive load. In addition, scrolling might have adverse effects on cognitive load, as it requires readers to continuously manage the amount of new text presented to them in order to continue reading, adding an additional task to the reading experience.

There is evidence from experimental research on different measures of cognitive load suggesting that reading on screen generates more cognitive load than reading on paper. Noyes et al. (2004) found that undergraduate participants reading on screen (1881-word text) scored significantly higher on the exerted effort subscale (d = 0.81) of the NASA Task Load Index. In a different study, students reported more stress (d = -0.55) and tiredness (d = -0.69) when reading five texts with an average of 1000 words on screen rather than on paper (Wästlund et al., 2005). In a study by Noyes and Garland (2003) on cognitive memory processes (see Tulving, 1985) reading from screen was found to impair learning processes.

Porion et al. (2016) could not replicate Noyes and Garland's (2003) findings with eight- and ninth grade students, attributing this to scrolling in the study by Noyes and Garland (2003), higher levels of computer familiarity and changes in print and display technology, specifically the shift from cathode ray tube (CRT) to liquid crystal display (LCD) monitors (Garland & Noyes, 2004; Noyes & Garland, 2008). Research into the effects of display settings in screen technology found that reducing the pixel density from 264 to 132 pixels per inch increased readers' level of discomfort (Mayr et al., 2017), indicating that improvements in display technology can affect the reading experience when reading from screen. However, there is still evidence that reading form screen has negative effects on eye strain (Köpper et al., 2016), especially when reading cognitive demanding texts (Rosenfield et al., 2015). The effects of modern LCD screens has only been the subject of Porion et al. (2016), who researched cognitive memory processes, while the studies on cognitive load have not been replicated. With screen inferiority also occurring with modern LCD monitors (Delgado et al., 2018), it is probable that the test medium can affect the reading experience in terms of cognitive load.

2.3 Test Mode: Fixed-item test (FIT) and Computer Adaptive Tests (CAT)

In addition to PPTs and CBTs, which are both considered fixed-item tests (FIT), as item selection is independent of examinees' responses, there are computer adaptive tests (CAT). The major difference between CBT and CAT is that CATs adapt subsequent item selection to students' prior responses in order to maximize the information gained from the test (Frey et al., 2017). The most informative items are neither too easy nor too difficult for the test-taker. Hence, adaptive tests avoid using items that are far above or below a given examinees' ability level. Because adaptive tests rarely administer items providing little information, they tend to be considerably more efficient than classical fixed-item tests (Davey, 2011; Weiss, 1982). Older studies often compared CATs with PPTs and found no significant differences in test scores (Alkhadher et al., 1998; Bergstrom, 1992). There are fewer studies comparing CATs with CBTs, but these also found no differences in test performance in mathematics (Ling et al., 2017; Martin & Lazendic, 2018) as well as reasoning (Ortner et al., 2014). Even fewer studies have compared CATs with FITs at the elementary school level. Kingsbury (2002) found no differences in reading comprehension between a PPT and a CAT in a large sample (N = 8560) of fourth and fifth grade students but did find significant differences in performance on the language use and mathematics portions of the test.

However, the test experience may not be equivalent between fixed-item and adaptive tests, because the latter are calibrated such that each student will answer only around half the items correctly on average, regardless of their ability. This can affect examinees' test experience, as their experienced performance on the adaptive test will not match their expected performance in FITs (Colwell, 2013; Ortner et al., 2014; Weiss, 1982). High-ability students will expect to answer many items correctly based on their experience with fixed-item tests, but item difficulty on the adaptive test will be calibrated to match students' ability. Thus, adaptive tests can result in lower levels of perceived probability of success and higher levels of situational fear of failure (Ortner et al., 2014), and even lower performance for students with higher levels of test anxiety (Lu et al., 2016; Ortner & Caspers, 2011). Conversely, low-ability students can find themselves answering more items correctly than they expected, potentially increasing their motivation and engagement (Betz, 1977; Weiss & Betz, 1973; see also Ling et al., 2017). In addition, the difference between an examinees' ability level and item difficulty of presented items will systematically be larger in the beginning of a CAT than in the end. Hence, items presented at the beginning of a CAT are more likely to be either considerably easier or harder than at a later stage of the CAT relative to an examinee's ability,

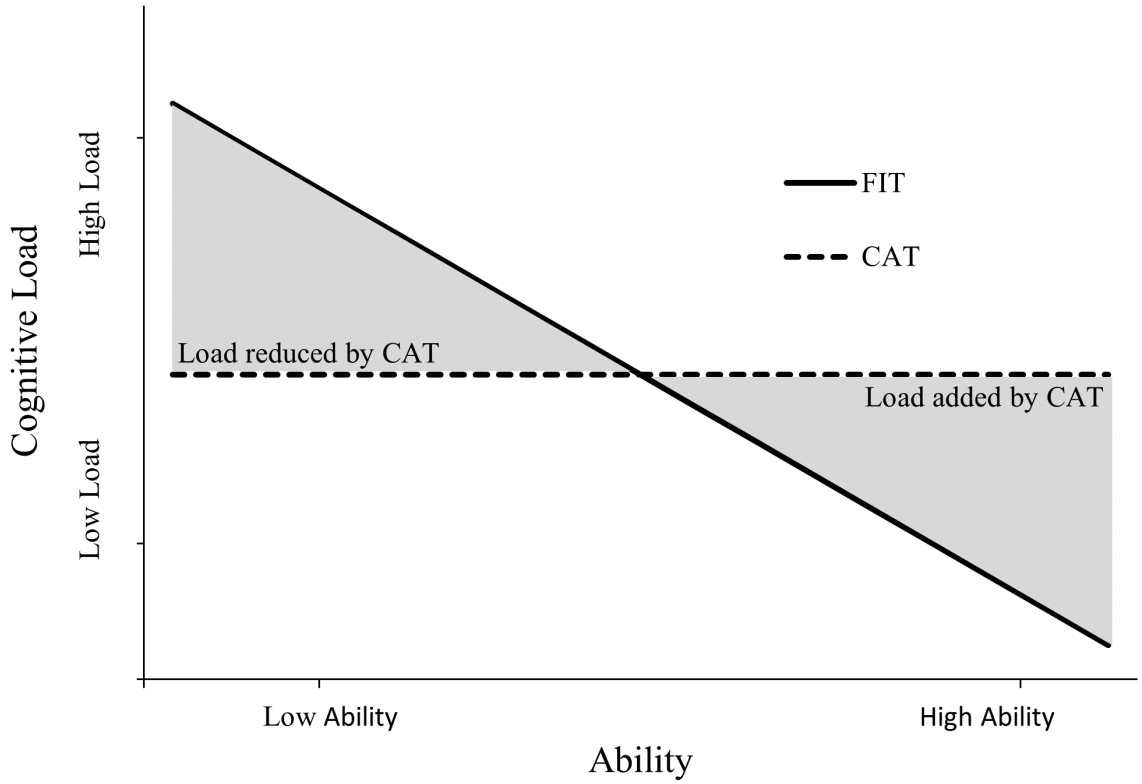
where items will be more challenging, since item difficulty aligns more with their ability. This change in challenge for examinees over the course of a CAT is systematic.

2.4 Cognitive Load in Fixed-Item and Adaptive Tests

Since intrinsic and extraneous cognitive load increase with item difficulty (Gupta & Zheng, 2020), CATs adapting the difficulty towards the examinee might affect examinees' experienced cognitive load differently than if they had taken a FIT. In a FIT, students above a certain ability level will always see some number of items considerably below or above their ability, which require fewer or more cognitive resources to solve. These items would quickly be filtered out in a CAT, as they offer little information once the test has been adjusted to the examinee's ability. Therefore, CATs will administer more items that are in difficulty close to the ability of the examinee and fewer items that differ strongly in difficulty from the ability of examinee. Items that are close in difficulty to the ability of an examinee are more challenging to solve than items that are lower in difficulty, thus generating intrinsic load. However, for low ability students, the adaptive test may reduce the experienced cognitive load, as items that are far too difficult will not be administered, while students of average ability would be equally as challenged by a CAT as by a FIT. Within this view, it is possible that an increase in cognitive load for high ability students is offset by the decrease in low ability students, resulting in equal average cognitive load on the test level. Figure 1 displays this schematic model on how much cognitive load is added or reduced within an adaptive test in comparison to a FIT.

Figure 1

Schematic model of the cognitive load generated by the difficulty of a FIT and CAT as a function of the examinee's ability



While there is considerable research on text experience in CATs with regard to test motivation (Weiss, 1982), test satisfaction (Ortner et al., 2013), or test anxiety (Ling et al., 2017), very little research has been conducted on how CATs affect cognitive load. To our knowledge, only Hayden (2005) measured cognitive load in an adaptive language test and argued that high-performing students reported higher levels of cognitive load descriptively, though they did not report the level of statistical significance. More research in how the adaptive nature of a CAT affects the experienced cognitive load is needed.

In summary, distinguishing between the test media of screen and paper and the test modes of adaptive and fixed tests leaves three different test formats for reading comprehension tests with potentially different test experiences. Unlike PPTs, CBTs and CATs are administered on screens, whereas PPTs and CBTs share the fixed mode of administration.

2.5 Purpose of the Present Study

The present study investigated the effects of different test formats on cognitive load during a reading comprehension test among fourth grade students. How different test formats affect readers' cognitive load has not been exhaustively researched, and no previous studies have investigated these effects in test situations in elementary school. Hence, this study sought to answer the following research question:

> How does cognitive load in a reading comprehension test in the fourth grade differ between paper-based (PPT), computer-based (CBA), and computer adaptive (CAT) administration methods?

The following differences between the test formats of PPT, CBT, and CAT are expected: based on the theoretical notion that cognitive load is negatively related to performance (Gupta & Zheng, 2020), differences in perception when reading on screen versus on paper (Mangen & Kuiken, 2014), and empirical findings related to screen inferiority (Delgado et al., 2018; Kong et al., 2018), our first hypothesis (H1) proposes that cognitive load is higher when reading on screen than on paper. Since CATs administer on average more items close to an examinees' ability level, fewer items considerably above or below their ability level are administrated. Within the context of the model in Figure 1, hypothesis 2 states that cognitive load can be described as a function of reading comprehension in the FIT, but not in the CAT.

In accordance with the working memory resource depletion aspect of cognitive load theory (Chen et al., 2018), it is further hypothesized that (H3) cognitive load increases over the course of a test in all test formats as more cognitive resources are depleted. Lastly, hypothesis 4 states that, because CATs take a number of items to calibrate to the examinee's ability level, it is possible that the increase in cognitive load over the course of a test as expected in H3 is higher in the CAT, as the items will be more challenging at the end of a test than at the beginning.

3. Method

3.1 Participants

The participants were $N = 212$ (49.50% female) fourth grade elementary school students attending 14 classes in eight schools in Western Germany. The students were on average 9.44 ($SD = 0.59$) years old, and 11.30% spoke a language other than German at home,

which is close to the national average (Mullis et al., 2017). University ethics approval was granted (GEKTUDO_2020_26), and only students with parental consent were included in the analyses. Students were assigned randomly to one of three experimental groups (N_{PPT} = 65, N_{CBT} = 69, N_{CAT} = 78). Simple random assignment on a per-student basis and independent of previous assignments was used to assign students to the groups. Participants in the conditions did not differ regarding gender, $\chi^2(2) = 1.12$, $p = .570$, country of origin, $\chi^2(2) = 4.00$, $p = .135$, computer self-efficacy, $F(2, 195) = 0.73$, $p = .482$ or academic achievement in German based on first-semester grades, $\chi^2(8) = 5.71$, $p = .680$.

3.2 Instruments

3.2.1 Reading Comprehension

To test reading comprehension, the German reading comprehension test FALKE was administered (Ludewig et al., 2021). This test has a calibrated item pool of 132 items based on 85 short texts with a mean length of 60.15 (SD = 17.67) words. The text pool contains 44 narrative and 41 expository texts. Texts differ strongly in the topic they cover and include topics such as music practice, refrigerators, the moon landing, the cinema, or food. The item pool comprises 69 text-based and 63 inference-based items. Each text is presented with a multiple-choice question with four response options, only one of which is correct. Some items share a text, but a given text is never administered more than once to the same person. The mean readability index of the texts was 31.40 (SD = 7.48) in the German version of the LIX (Bamberger & Vanacek, 1984), which is appropriate for children's and youth literature (Lenhard, 2019). The item pool was used to build the three test formats.

The test design and layout in each test format was made to be as comparable as possible. All test formats had a fixed test length of 25 items and the same content-balancing sequence. The computer screens (Lenovo Miix 320-10ICR with 1280 x 800 resolution at 60 Hz on a 10.1-inch display) used for the CBT and CAT had nearly the same size as the paper used in the PPT (ISO 216 A5 in landscape format; 10.08-inch diagonal); additionally, the font type (Arial), font size (text: 15 pt. or 20 pixels; answers: 13.5 pt. or 18 pixels) and item layouts were designed to be very similar. Students completing the PPT were instructed not to return to previous pages, while the ability to return to previous questions was disabled in the CBT and CAT.

For the FIT (PPT and CBT), 25 items were selected such that the amount of provided information for the expected ability distribution of fourth graders was maximized. Items were

still content balanced regarding text- and question types. The selected items were administered in six booklets with a nested Latin-square design to balance possible context, content, and order effects (Frey et al., 2009).

The CAT (1) began with an item of average difficulty, (2) then selected items based on the maximum Fisher information, while keeping the content balanced (equal proportions of narrative texts and expository texts as well as text-based and inference-based items presented), (3) and then stopped after 25 items. The CAT used expected a posteriori (EAP) estimates for the preliminary ability estimates (Bock & Mislevy, 1982). Students' final scores were estimated in all three test formats based on weighted likelihood estimates (Warm, 1989) with fixed calibrated item parameters using the R (R Core Team, 1999) package catR (Magis & Raîche, 2012) in order to make the final scores in the three test formats comparable. WLE reliability was good in all test formats: PPT (Rel_{WLE} = .83), CBT (Rel_{WLE} = .77), and CAT (Rel_{WLE} = .88). FITs and CAT did not differ in difficulty on the test level, $t(210) = 1.39$, p = .168.

In the PPT, not-reached items were considered as not administered. The CBT and CAT used forced-choice response formats. To reduce the influence of rapid guessing behavior, we used an item response time threshold of 7.5 seconds; responses in the CBT and CAT conditions were considered as not administered if response times were faster than this threshold (for a detailed explanation of this process, see Wise and DeMars, 2006). This applied to 168 (4.95%) valid responses in the CBT and CAT conditions.

3.2.2 Cognitive Load

Cognitive load was measured with a self-report measure. Self-report measures for cognitive load have been successfully applied to children (e.g. Chu, 2014; van de Weijer-Bergsma & van der Ven, 2021). The cognitive load measure was based on an adaption of the naïve rating scale by Klepsch et al. (2017) for children (Altmeyer et al., 2022). The scale is based on the cognitive load theory and consists of five items measuring intrinsic cognitive load with two items and extraneous cognitive load with three items on a four-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (4). The items were adapted for children and to the reading task at hand. This resulted in a 5-item cognitive load short scale. Intrinsic load was measured with the items "In the reading tasks, many things needed to be considered at the same time" ["Bei den Leseaufgaben musste man viele Dinge gleichzeitig beachten"] and "the reading tasks were complicated" ["Die Leseaufgaben waren kompliziert"]. Extraneous load was measured with the items "The reading tasks were difficult,

because important information was not always easy to find" ["Die Leseaufgaben zu bearbeiten war schwierig, da wichtige Dinge nicht immer leicht zu finden waren."], "It was difficult to read the tasks, because important information was hidden" ["Es war schwierig die Aufgaben zu lesen, weil wichtige Informationen versteckt waren"], and "The manner in which the tasks were displayed made them difficult to understand" ["Die Art, wie die Aufgaben abgebildet waren, hat es schwierig gemacht, sie zu verstehen".].

3.2.3 Demographic Measures

Demographic variables regarding age, gender, country of origin, socioeconomic background and language spoken at home, as well as computer self-efficacy were collected. Language spoken at home was measured with a four-point scale ranging from "At home, I always speak German and never another language" to "At home, I often speak German and sometimes another language", "At home, I sometimes speak German and often another language", and finally "At home, I never speak German and always another language". Socioeconomic background was operationalized with the number of books at home (OECD, n.d.) and the highest level of education completed by either of the students' parents.

3.3 Design & Procedure

The study used a within- and between-subject design. First, a demographic questionnaire on students' age, gender (m/f/d), language spoken at home, country of origin, and socioeconomic background was administered. Afterwards, the students in each class were randomly assigned to one of three test formats: PPT, CBT, or CAT (between-subject). Differences between the CBT and the CAT were not explained to the students. In each test format, students were presented a 25-question reading comprehension test. At the midpoint (after the 12th item) and at the end of the test, students were given a brief questionnaire on their current cognitive load level (within-subject). Students had 40 minutes in total to complete the test and associated questionnaires, of which 35 minutes were allocated for the test and five minutes for the midpoint- and post-test cognitive load measure. Students that did not complete all 25 items within 35 minutes (N = 137; 64.62%) were instructed to complete the cognitive-load measure in the post-test. Most students (N = 174; 82.08%) answered at least 20 of the 25 items. There were no significant differences in the number of students who completed all 25 items in the PPT (60%), CBT (70%), or CAT (64%); $\chi^2(2, N = 212) = 1.35$, $p = .508$. Missing values on the cognitive load measure were imputed using the R package mice (van Buuren & Groothuis-Oudshoorn, 2011). In total, 10% of responses to the cognitive load items were

missing. As predictors, in addition to test format, test performance, and measurement point, demographic variables such as gender, country of birth, language spoken at home and socioeconomic status as measured by the number of books and the highest educational level achieved by either of both parents were used. Cognitive load was imputed using a linear mixed-effects model, while the predictor variables were imputed with predictive mean matching. Ten datasets, one for each percentage point of missing data in the cognitive load measure (White et al., 2011), with 20 repetitions were imputed.

3.4 Analysis

The students' responses regarding experienced cognitive load were modeled using a latent regression approach within a linear mixed-effects model framework (de Boeck et al., 2011). The analysis was fitted using the package lme4 (Bates et al., 2015) in the R environment (R Core Team, 1999).

We used a linear mixed-effects model rather than a repeated measures ANOVA due to its greater flexibility, better handling of missing data, and opportunity to account for measurement error in the outcome measure. The baseline model explains the experienced cognitive load rating via an item random effect and a person random effect. The item random effect represents the item difficulty, i.e., the average cognitive load rating of the item, while the person random effect represents the person's individual tendency to experience cognitive load. Additionally, the baseline model includes the fixed effect of time, i.e., the average increase in cognitive load from midway through the test to the end of the test, and the student's reading comprehension score (z-score). The first hypothesis (H1) was tested by adding the test medium (screen versus paper) and the interaction between medium and time to the model. H1 is supported if the medium explains experienced cognitive load over and above time and reading comprehension. The theoretical model of the relationship of test mode and performance (H2) was tested with a linear mixed effect model. Cognitive load was explained for the FIT and the CAT separately with the same item and person random effects and the reading comprehension score and the measurement point as fixed effects. The hypothesis is supported if cognitive load is explained by the reading comprehension score in the FIT, but not in the CAT. In addition, since differences in cognitive load between FITs and CATs would make the model untenable, the test mode (fixed-item test versus adaptive test) was added to the baseline model to see if there were differences in the cognitive load between test modes. The third hypothesis (H3) is supported if the measurement point explains experienced

cognitive load. Lastly, H4 specifically examined the differential growth of cognitive load in the test formats. H4 is therefore supported if the interaction between measurement point and test format explains experienced cognitive load.

4. Results

4.1 Descriptive Statistics

Table 1 shows descriptive statistics for the reading comprehension test scores and cognitive load measure for all conditions and at all measurement points. While test scores were descriptively highest in the CBT conditions, the differences were not statistically significant, $F(2, 209) = 1.57$, $p = .210$. A screen inferiority effect as described in Delgado et al. (2017) of $d_c = 0.21$ could not be detected, though the statistical power with the present sample size was only 23 %. Due to the theoretical relationship between cognitive load and performance (Gupta & Zheng, 2019) and the descriptive differences in test performance, models controlled for reading comprehension test scores.

A slight trend towards higher cognitive load over the course of the test can be observed. The WLE reliabilities for the reading comprehension test were good in all formats. Reliability for the total cognitive load scales were overall and for most subgroup measurements satisfactory (Taber, 2018), but poorer at the midway point in the CBT and CAT conditions. The reliabilities for the total load were comparable with previous applications of the scale for adults (e.g. .74 in Albus et al., 2021; .86 in Klepsch et al., 2017).

Table 1

Means (M), standard deviations (SD) and reliability coefficients for the dependent variable cognitive load and the independent variable reading comprehension

Test Format	*Reading comprehension test z-scores*			*Measurement point*	*Cognitive load*		
	M	*SD*	Rel_{wle}		*M*	*SD*	α
Paper	-0.07	0.89	.83	Midway	2.08	0.84	.85
				Post	2.15	0.92	.86
Computer	0.17	0.82	.77	Midway	1.90	0.56	.58
				Post	2.06	0.79	.83
Adaptive	-0.09	1.20	.88	Midway	1.90	0.62	.69
				Post	2.21	0.88	.84
Total	0.00	1.00		Midway	1.95	0.67	.74
				Post	2.14	0.86	.85

Note. $N = 212$, $N_{PPT} = 65$, $N_{CBT} = 69$, $N_{CAT} = 78$.

Table 2 shows the intercorrelations between the reading comprehension test z-scores and the cognitive load measures at the midway point and post-test. Cognitive load between the two measurement points was strongly correlated. Reading comprehension and cognitive load were negatively correlated with each other, as theoretically expected.

Table 2

Inter-correlations between cognitive load measures and reading comprehension

	Reading comprehension	Cognitive load (midway)
Reading comprehension	-	
Cognitive load (midway)	**-.15**	-
Cognitive load (post-test)	**-.18**	**.65**

Note. $N = 212$; correlations in bold are significant at $p < .05$.

4.2 How does cognitive load differ between PPT, CBT, and CAT?

Linear mixed-effects models were used to investigate the hypotheses. First, a baseline model was computed where cognitive load was predicted by a fixed effect of the measurement point of the cognitive load measure, the z-standardized reading comprehension ability estimate, and random effects for individual and item characteristics. The baseline model (M1) in Table 3 displays the results of the linear mixed-effects model.

Statistically significant effects of measurement point and reading comprehension test performance on cognitive load were found. Cognitive load increased between the first and second measurement point, indicating that in accordance with H3, students experienced more cognitive load at the end of the test than in the middle. Test performance was negatively associated with cognitive load, as higher-performing students reported less cognitive load. Two linear mixed-effects models were used to investigate differences between the test formats by comparing the entire sample either across media (M2: PPT versus CBT and CAT) or across modes (M3: PPT and CBT versus CAT). Lastly, a fourth model (M4) was calculated in which both interactions described in M2 and M3 were included simultaneously. Table 3 displays these alternative models alongside the baseline model. In all alternative models, the main effects of both measurement point and reading comprehension were significant predictors of cognitive load.

Table 3

Regression explaining cognitive load controlling for reading comprehension, measurement point, and test format

	M1: Baseline		*M2: Paper vs Screen*		*M3: FIT vs CAT*		*M4: Full model*	
	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Fixed Effects								
(Intercept)	**2.04**	0.07	**2.06**	0.07	**2.04**	0.07	**2.06**	0.08
RC (z-score)	**-0.13**	0.05	**-0.12**	0.05	**-0.13**	0.05	**-0.12**	0.05
MP (post = 1)	**0.09**	0.02	**0.08**	0.02	**0.10**	0.02	**0.09**	0.02
Medium (screen = 1)			-0.04	0.05			-0.05	0.06
Mode (adaptive = 1)					-0.01	0.05	0.01	0.06
MP x Medium (screen)			0.03	0.02			0.02	0.02
MP x Mode (adaptive)					**0.04**	0.02	0.03	0.02
Random Variances								
σ^2Person	0.409		0.409		0.411		0.411	
σ^2Item	0.011		0.011		0.011		0.011	
σ^2Residual	0.589		0.588		0.588		0.588	

Note. RC = Reading comprehension (z-score), MP = Measurement point; coefficients in bold are significant at $p < .05$; sample size: $N = 212$, items: $I = 5$, measurement points: 2, observations = 2120 (212 x 5 x 2).

There was no main effect of test medium on cognitive load, suggesting that students do not experience more cognitive load when taking a reading comprehension test on screen; thus, H1 should be rejected. In fact, at the midway point of the test, cognitive load was

descriptively higher in the PPT condition than in both screen conditions. In both M3 and M4, no main effect of administration was found. Model 4 also indicated no main effect of test medium or test mode on cognitive load.

4.3 To What Extent Does Test Performance Affect Cognitive Load in FITs and CATs?

For the differences between FITs and CATs regarding the effects of reading comprehension (H2), model 5 explained cognitive load using reading comprehension and measurement point as fixed effects found for the FITs and model 6 for the CAT. Table 4 displays these models. Reading comprehension was a statistically significant predictor for cognitive load only in model 5 for the FIT condition.

Table 4

Regression explaining cognitive load for the FIT and the CAT controlling for reading comprehension and measurement point

	M5: FIT		*M6: CAT*	
	β	*SE*	β	*SE*
Fixed Effects				
(Intercept)	**2.05**	0.09	**2.04**	0.08
Reading comprehension (z-score)	**-0.15**	0.06	-0.10	0.08
Measurement point (post = 1)	**0.06**	0.02	**0.14**	0.03
Random Variances				
σ^2Person	0.438		0.365	
σ^2Item	0.018		0.001	
σ^2Residual	0.541		0.666	

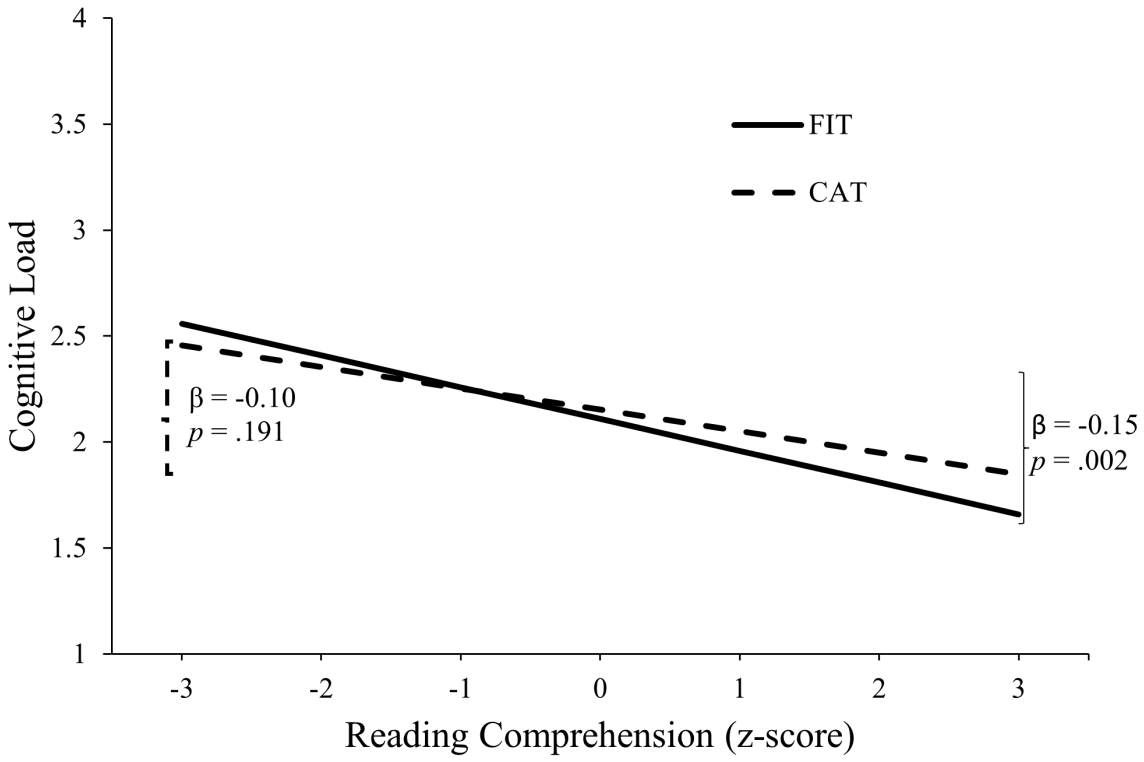
Note. Coefficients in bold are significant at $p < .05$; sample size: $N = 212$, items: $I = 5$, measurement points: 2, observations = 2120 (212 x 5 x 2).

The regression lines are shown in Figure 2. In the FIT condition, reading comprehension was a significant negative predictor for cognitive load. As hypothesized, lower levels of performance were associated with higher levels of cognitive load. This relationship was not found in the CAT, but a test for the equality of coefficients found no difference

between the coefficients for reading comprehension, $z = 1.00$, $p = .159$ (Paternoster et al., 1998).

Figure 2

Reading comprehension test z-scores and post-test cognitive load for the CAT and the FIT



Note. The coefficient for the measurement point was set to 1.

4.4 To What Extent Does Cognitive Load Change Over the Course of the Test?

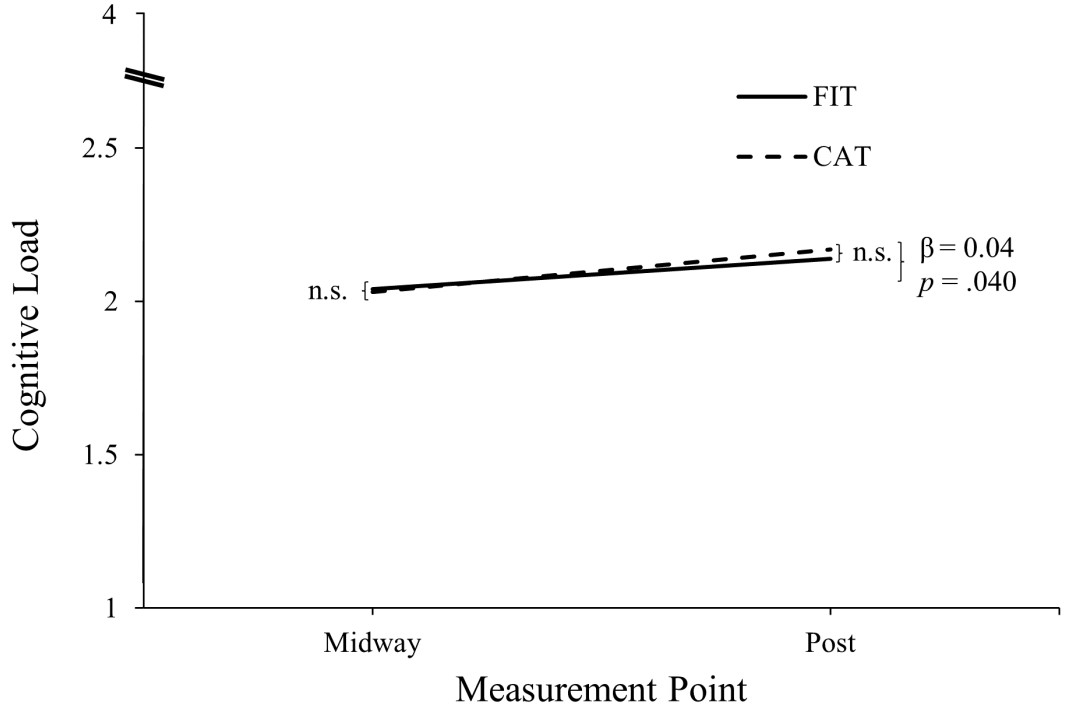
As previously mentioned, it was expected that cognitive load would increase over the course of the test (H3) and would increase at different rates between the test formats (H4). Differential increases in cognitive load between the test formats were measured via the interaction effects in Models 2, 3, and 4. The linear mixed-effects model provided strong support for H3 in that measurement point was a statistically significant predictor of cognitive load in all models. In M2, cognitive load did not increase more in the screen condition than in the paper condition, but the interaction effect between measurement point and test mode reached significance in M3, suggesting that students' experienced cognitive load increases more when taking a CAT rather than a FIT. This is supported by M5 and M6, where measurement point was a significant predictor for cognitive load in both the FIT and the CAT, though the coefficient was significantly larger in the CAT, $z = 2.22$, $p = .013$. However, in

M4, which included all possible predictors, this interaction effect did not reach statistical significance. Instead, no main or interaction effect of test format reached significance in M4.

A graphical illustration of the interaction effect in M3 can be found in Figure 3. The figure shows that cognitive load did not differ significantly between the FITs and CAT at the midway or end point of the test. The significant effect is displayed as the increase in the growth rate of cognitive load over time, which is higher in the CAT format.

Figure 3

Figure displaying the interaction effect between test mode and time



5. Discussion

This study examined how fourth grade students experienced cognitive load over the course of a reading comprehension test administered in three different test formats. This study found no differences in experienced cognitive load between reading comprehension tests administered on paper, on the computer, or in a computer adaptive format. Students experienced more cognitive load at the end of each test than in the middle. This intuitive effect offers further support for the working memory resource depletion effect (Chen et al., 2018), as it can be reasoned that working memory resources deplete over the course of a test, meaning that more mental effort is required to complete test items. The increase in cognitive load did

not differ between test media (computer versus pencil-and-paper), but when comparing the adaptive test format with the fixed-item test format, there was limited evidence for the tendency that cognitive load may increase more in an adaptive test, though the effect was very small and not consistent across models.

Furthermore, students who performed well on the reading comprehension test reported lower levels of cognitive load in the FIT, but not in the CAT. For the FIT, this finding is in line with previous studies finding correlations between high levels of cognitive load and lower levels of test performance (Mayes et al., 2001; van de Weijer-Bergsma & van der Ven, 2021). The direction of this relationship is likely reciprocal. High-ability examinees are likely to require fewer resources to solve problems, reducing their experienced cognitive load; on the other hand, cognitive load can impair the problem-solving process, decreasing performance (Sweller et al., 2019).

5.1 Screen Effects

Based on cognitive load theory and screen inferiority, it was hypothesized that reading on screen would lead to more cognitive load than reading from paper. This assumption was not supported by the data, as cognitive load did not differ significantly in the PPT and CBT conditions. A screen inferiority effect could not be replicated with this sample. Several factors might explain the present study's incongruity with previous research (Delgado et al., 2018; Kong et al., 2018). This study placed a core focus on reducing the differences between reading on screen and reading on paper to the actual differences in display. Additional aspects, such as differences in presentation format (i.e., ISO 216 A4 portrait format versus widescreen monitor) or scrolling, were consciously eliminated in this study. These aspects are suspected to contribute to screen inferiority (Delgado et al., 2018). In addition, similarly to the study by Porion et al. (2016), modern display technologies might reduce the impact of screens on examinees' cognitive load. Furthermore, while computer familiarity does not seem to affect performance (Aesaert & van Braak, 2015; Halamish & Elbaz, 2020), performance may still be affected by the cognitive demands of interacting with computers, such as the input method that differs between pencil-and-paper and mouse-and-keyboard. The majority of today's students have substantial experience working with computers, which might reduce load imposed by interacting with a computer in a different and unfamiliar way. Lastly, the study was conducted during a gap between school closures due to the COVID-19 pandemic. The students in this sample spent a considerable amount of the school year in distance learning,

increasing their exposure to digital media for learning purposes in comparison to previous studies.

5.2 Test Mode Effects

The results showed that the experienced cognitive load could be explained with the reading comprehension test scores in the FIT, but in the CAT test performance was not a statistically significant predictor of cognitive load. The coefficient in the CAT was smaller but still negative and the difference between the coefficients was not significant. The results suggest a tendency that in CATs, cognitive load is more evenly distributed among high and low ability students than in the FIT, which was proposed in H2. Since there was no difference in cognitive load between FITs and CATs, the results can be interpreted as supportive for the schematic model that CATs increase the cognitive load of high ability examinees and reduce the load of low ability examinees, though more research with larger samples is needed to make conclusive statements on H2.

There was some evidence for a tendency that cognitive load increased more for students who completed the adaptive test. Though the effect was small and not consistent across models, this is theoretically plausible, since one aspect of adaptive testing is that item difficulty becomes successively better aligned with each examinee's abilities as more items are administered (Davey, 2011; Weiss, 1982). Therefore, the difference in cognitive load between fixed and adaptive tests might only arise late in a given test-taking period. Over a sufficiently long testing period, adaptive reading comprehension tests may induce higher levels of cognitive load among fourth grade students than their FIT counterparts.

This finding contextualizes the efficiency advantage of CATs over FITs. CATs can reach a higher level of measurement precision after the same number of administered items (Davey, 2011). Therefore, CATs can be more time-efficient (i.e. precision per time). However, this increased precision goes along with a higher increase or "cost" of cognitive load. Thus, the cognitive load efficiency (i.e. precision per cognitive effort) advantage of CATs over FITs might be smaller than the time efficiency advantage. This suggests that achieving a certain amount of measurement precision necessitates a similar amount of cognitive effort irrespective of the test mode.

5.3 Limitations and Strengths

This study was conducted in between periods of distance learning due to the COVID-19 pandemic. A number of limitations resulting from this unique circumstance must be mentioned. Due to reduced class sizes, the sample size in this study was limited. This also limited the power so that effect sizes of $d_c > 0.43$ could be identified with a power of 80 %. However, effect sizes related to the difference in display medium, such as screen inferiority, can be smaller. Furthermore, students' computer familiarity and usage of computers as learning tools might differ from previous studies due to the widespread adoption of digital media-supported distance learning.

Despite this, due to the within-class experimental design, the differences in cognitive load by test mode and medium can be compared, and the effects should be interpretable. In this study, we applied an experimental design to compare PPTs and CBTs with regard to their test medium (screen versus paper). In many use cases, test developers or teachers will make design choices that are specific to a given test medium, such as allowing for scrolling or hyperlinks. This study makes no statements regarding these aspects.

6. Conclusion

In this study, we looked at the differences between three different test formats (PPT, CBT, CAT) in experienced cognitive load during and after a reading comprehension test among fourth grade students. Cognitive load increased over the course of the test, lending support to the view that working memory is a limited resource that is depleted during task performance. Reading comprehension test scores and cognitive load were negatively associated in the FIT. There were no differences in cognitive load between reading on screen and reading on paper. Screen reading does not seem to induce any additional cognitive load in short texts during elementary school. Since the finding of screen inferiority could not be replicated in the present sample, no statements about a possible relationship between screen inferiority and cognitive load can be made. Instead, this study's results suggest that using computers to assess fourth-grade students' reading comprehension via short texts has no detrimental effects on either experienced cognitive load or performance. The same can be said of CATs, provided they are administered in an efficient manner. In order to generalize these findings, more research into the effects of scrolling on readers' perception could help to determine appropriate text lengths, which tend to increase as students reach higher grades in school, for computer-based test formats. This would not only advance research on the

cognitive effects of the test formats, but also increase our understanding of how screen inferiority emerges.

References

Aesaert, K., & van Braak, J. (2015). Gender and socioeconomic related differences in performance based ICT competences. *Computers & Education*, *84*, 8–25. https://www.sciencedirect.com/science/article/pii/S0360131515000020

Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, 166, 104154.

Alkhadher, O., Clarke, D. D., & Anderson, N. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the Differential Aptitude Tests. *Journal of Occupational and Organizational Psychology*, *71*(3), 205–217. https://doi.org/10.1111/j.2044-8325.1998.tb00673.x

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37. https://doi.org/10.18178/ijiet.2018.8.1.1008

Altmeyer, K., Barz, M., Lauer, L., Peschel, M., Sonntag, D., Brünken, R., & Malone, S. (2022). *Digital ink and differentiated subjective ratings for cognitive load measurement in middle childhood* [Manuscript submitted for publication]. Department of Education, Saarland University.

Ball, L. J., Threadgold, E., Solowiej, A., & Marsh, J. E. (2018). Can intrinsic and extrinsic metacognitive cues shield against distraction in problem solving? *Journal of Cognition*, *1*(1), 15. https://doi.org/10.5334/joc.9

Bamberger, R., & Vanacek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben* [Reading-Unterstanding-Learning-Writing]. Jugend und Volk.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, M., McElvany, N., & Kortenbruck, M. (2010). Intrinsic and extrinsic reading motivation as predictors of reading literacy: A longitudinal study. *Journal of Educational Psychology, 102*(4), 773–785. https://doi.org/10.1037/a0020084

Bergstrom, B. A. (1992). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Fransico, CA. Retrieved September 6, 2022, from https://eric.ed.gov/?id=ED377228

Betz, N. E. (1977). Effects of immediate knowledge of results and adaptive testing on ability test performance. *Applied Psychological Measurement*, *1*(2), 259-266.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1-9.

Chall, J. S. (1983). Literacy: Trends and explanations. *Educational Researcher*, *12*(9), 3-8.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, *30*(2), 483–501. https://doi.org/10.1007/s10648-017-9426-2

Chu, H. C. (2014). Potential negative effects of mobile learning on students' learning achievement and cognitive load—A format assessment perspective. *Journal of Educational Technology & Society, 17(1), 332-344.*

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies*, *1*(2), 50–60. http://redfame.com/journal/index.php/jets/article/view/101

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/s0022-5371(80)90312-6

Davey, T. (2011). A guide to computer adaptive testing systems. *Council of Chief State School Officers*. Retrieved September 6, 2022, https://eric.ed.gov/?id=ed543317

Dawidowsky, K., Holz, H., Schwerter, J., Pieronczyk, I., & Meurers, D. (2021). Development and evaluation of a tablet-based reading fluency test for primary school children. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM. https://doi.org/10.1145/3447526.3472033

de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12). https://doi.org/10.18637/jss.v039.i12

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. https://doi.org/10.1016/j.edurev.2018.09.003

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Dealing with item position effects when developing computer adaptive tests]. *Diagnostica*, *63*(3), 167–178. https://doi.org/10.1026/0012-1924/a000173

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests [Effects of adaptive testing on test-taking motivation with the example of the Frankfurt Adaptive Concentration Test]. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Garland, K. J., & Noyes, J. M. (2004). Crt monitors: Do they interfere with learning? *Behaviour & Information Technology*, *23*(1), 43–52. https://doi.org/10.1080/01449290310001638504Gupta, U., & Zheng, R. Z. (2020). Cognitive Load in Solving Mathematics Problems: Validating the Role of Motivation and the Interaction among Prior Knowledge, Worked Examples, and Task Difficulty. *European Journal of STEM Education*, *5*(1), 5.

Halamish, V., & Elbaz, E. (2020). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education*, *145*, 103737. https://doi.org/10.1016/j.compedu.2019.103737

Hayden, J. J. (2005). Breaking the camel's back: Cognitive load and reading Chinese. *In International Interdisciplinary Conference on Hanzi renzhi – How Western learners discover the world of written Chinese.* Retrieved September 6, 2022, from https://chinesisch.fb06.uni-mainz.de/files/2018/11/hanzirenzhi_papers_hayden.pdf

Helm, C., Huber, S., & Loisinger, T. (2021). Was wissen wir über schulische Lehr-Lern-Prozesse im Distanzunterricht während der Corona-Pandemie? - Evidenz aus Deutschland, Österreich und der Schweiz [What do we know about learning and teaching processes in remote learning during the corona-pandemic? Evidence from Germany, Austria, and Switzerland]. *Zeitschrift für Erziehungswissenschaft, 24(2),* 237-311.

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [IGLU 2016. International comparison of reading competencies of elementary school students in Germany]. Waxmann. https://www.pedocs.de/frontdoor.php?source_opus=15476

Kingsbury, G. G. (2002). An empirical comparison of achievement level estimates from adaptive tests and paper-and-pencil tests. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, *123*, 138–149. https://www.sciencedirect.com/science/article/pii/S0360131518301052

Köpper, M., Mayr, S., & Buchner, A. (2016). Reading from computer screen versus reading from paper: does it still make a difference? *Ergonomics*, *59*(5), 615-632.

Leahy, W., & Sweller, J. (2019). Cognitive load theory, resource depletion and the delayed testing effect. *Educational Psychology Review*, *31*(2), 457–478. https://doi.org/10.1007/s10648-019-09476-2

Lenhard, W. (2019). *Leseverständnis und Lesekompetenz: Grundlagen - Diagnostik – Förderung* [Reading comprehension and reading competence: basics – diagnostics – development]. Stuttgart: Kohlhammer Verlag.

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P., & van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511.

Lu, H., Hu, Y., Gao, J., & Kinshuk (2016). The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing. *Computers & Education*, *100*, 45–55. https://doi.org/10.1016/j.compedu.2016.04.012

Ludewig, U., Trendtel, M., Schlitter, T., & McElvany, N. (2021). Adaptives Testen von Textverständnis in der Grundschule. *Diagnostica, 68*(1), 39-50.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8). https://doi.org/10.18637/jss.v048.i08

Mangen, A., & Kuiken, D. (2014). Lost in an iPad. *Scientific Study of Literature*, *4*(2), 150–177. https://doi.org/10.1075/ssol.4.2.02man

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68. https://www.sciencedirect.com/science/article/pii/S0883035512001127

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27. https://psycnet.apa.org/record/2017-17470-001

Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, *28*(6), 367–378. https://www.sciencedirect.com/science/article/pii/S0169814101000439

Mayr, S., Köpper, M., & Buchner, A. (2017). Effects of high pixel density on reading comprehension, proofreading performance, mood state, and physical discomfort. *Displays*, *48*, 41–49. https://doi.org/10.1016/j.displa.2017.03.002

McDougall, J., & Potter, J. (2019). Digital media learning in the third space. *Media Practice and Education*, *20*(1), 1–11. https://doi.org/10.1080/25741136.2018.1511362

Mullis, I. V. S., Martin, M. O., & Hooper, M. (2017). Measuring changing educational contexts in a changing world: Evolution of the TIMSS and PIRLS questionnaires. In *Cognitive Abilities and Educational Outcomes* (pp. 207–222). Springer, Cham. https://doi.org/10.1007/978-3-319-43473-5_11

National Center for Education Statistics. (2019). *The nation's report card. National Assessment of Eduucational Progress*. Retrieved September 9, 2022, from https://www.nationsreportcard.gov/reading?grade=4

Noyes, J. M., & Garland, K. J. (2003). VDT versus paper-based text: Reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, *31*(6), 411–423. https://www.sciencedirect.com/science/article/pii/S0169814103000271

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375.

Noyes, J. M., Garland, K. J., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect? *British Journal of Educational Technology*, *35*(1), 111–113.

OECD. (n.d.). *Student questionnaire for PISA 2015*. Retrieved September 6, 2022, from https://www.oecd.org/pisa/data/2015database/

OECD. (2019). *PISA 2018 results: Vol. I. What students know and can do*. OECD Publishing Press, Paris. https://doi.org/10.1787/5f07c754-en.

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*. https://econtent.hogrefe.com/doi/full/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E., & Gerstenberg, F. X. R. (2013). Skilled but unaware of it: Cat undermines a test taker's metacognitive competence. *European Journal of Psychology of Education*, *28*(1), 37–51. https://doi.org/10.1007/s10212-011-0100-7

Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I will probably fail. *European Journal of Psychological Assessment*. https://econtent.hogrefe.com/doi/full/10.1027/1015-5759/a000168

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*(4), 859-866

Piolat, A., Roussey, J. Y., & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, *47*(4), 565–589. https://doi.org/10.1006/ijhc.1997.0145

Porion, A., Aparicio, X., Megalakaki, O., Robert, A., & Baccino, T. (2016). The impact of paper-based versus computerized presentation on text comprehension and memorization. *Computers in Human Behavior*, *54*, 569–576. https://doi.org/10.1016/j.chb.2015.08.002

Potter, J., & McDougall, J. (2017). *Digital media, culture and education: Theorising third space literacies*. London: Palgrave Macmillan.

R Core Team. (1999). *Writing R extensions*. Retrieved September 9, 2022, from https://mirrors.nju.edu.cn/cran/doc/manuals/r-devel/r-exts.pdf

Rosenfield, M., Jahan, S., Nunez, K., & Chan, K. (2015). Cognitive demand, digital screens and blink rate. *Computers in Human Behavior*, *51*, 403-406.

Seigneuric, A., Ehrlich, M. F., Oakhill, J. V., & Yuill, N. M. (2000). Working memory resources and children's reading comprehension. *Reading and Writing*, *13*(1/2), 81–103. https://doi.org/10.1023/A:1008088230941

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education*, *151*, 103861. https://doi.org/10.1016/j.compedu.2020.103861

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van de Weijer-Bergsma, E., & van der Ven, S. H. (2021). Why and for whom does personalizing math problems enhance performance? Testing the mediation of enjoyment and cognitive load at different ability levels. *Learning and Individual Differences*, *87*, 101982. https://doi.org/10.1016/j.lindif.2021.101982

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, *21*(2), 377–394. https://www.sciencedirect.com/science/article/pii/S0747563204000202

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or Adaptive?* (Research Report 73-1). Minneapolis: Department of Psychology, University of Minneapolis.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Wigfield, A., Gladstone, J., & Turci, L. (2016). Beyond cognition: Reading motivation and reading comprehension. *Child Development Perspectives*, *10*(3), 190–195. https://doi.org/10.1111/cdep.12184

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Oecd Education Working Papers* (Vol. 209). OECD. Retrieved September 21, 2022, from https://www.oecd-ilibrary.org/education/introduction-of-multistage-adaptive-testing-design-in-pisa-2018_b9435d4b-en https://doi.org/10.1787/b9435d4b-en

3.3 Beitrag III: Effects of Test Mode and Medium on Elementary School Students‘ Test Experience

Brüggemann, T., Ludewig, U., Lorenz, R., & McElvany, N. (2023). Effects of Test Mode and Medium on Elementary School Students‘ Test Experience. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000767

Dieses Kapitel enthält keine Kopie des abgedruckten Artikels und weicht von der publizierten Version leicht ab.

Zusammenfassung

Der Einsatz von digitalen Medien im Bildungsbereich nimmt stetig zu und birgt große Potentiale. Beim Testen können neben computerbasierten Tests auch adaptive Testformate genutzt werden, die sich an die Testpersonen anpassen. Die Unterschiede zwischen diesen Testformaten können sich allerdings auf die Vergleichbarkeit der Tests auswirken. Papierbasierte Tests (PPT) unterschieden sich von computerbasierten (CBT) und computeradaptiven Tests (CAT) im Darstellungsmedium und adaptive Tests von fixierten Tests im Testmodus. In dieser Studie wird untersucht, wie sich diese Testformate auf die Zustandslesemotivation und Zustandsangst von 387 (46.3 % weiblich) Schüler*innen der vierten Klasse in einem standardisierten Lesekompetenztest auswirken. Unter Kontrolle von allgemeiner Testängstlichkeit und der Zustandsangst vor dem Test zeigte eine Ko-Varianzanalyse mit Messwiederholung keine Unterschiede zwischen den Testformaten in der Zustandsangst in der Mitte und am Ende des Tests. Eine Ko-Varianzanalyse mit Messwiederholung für die Zustandslesemotivation unter Kontrolle der allgemeinen Lesemotivation fand jedoch eine signifikant erhöhte Zustandslesemotivation für Schüler*innen, die den Test am Computer bearbeiteten (CBT und CAT), in der Mitte des Tests. Dieser Unterschied wurde am Ende des Tests nicht gefunden, da die Zustandslesemotivation über den Testverlauf im CBT und CAT auf das Level des PPTs absank.

Abstract

The use of digital media in education can bring great benefits and its use in schooling is steadily increasing. Administrating paper- versus computer-based as well as fixed-item versus adaptive tests could create differences in test experience, which can threaten the comparability of test results. This study investigated how the pen-and-paper, computer-based, and computer adaptive test formats of a standardized reading comprehension test affect test anxiety and motivation among German fourth grade students. A within-class randomized field trial with 387 fourth graders (aged 9 to 10; 46.3% female) was conducted. Repeated measures ANCOVA revealed no differences in state test anxiety between the test formats when controlling for trait test anxiety and pre-test state anxiety, but state reading motivation was initially higher when reading on a screen, controlling for trait reading motivation. However, this difference diminishes over the course of the test. Implications for using digital media in elementary school test situations are discussed.

Introduction

Digital media offer opportunities for teachers to improve the teaching, monitoring and evaluation of their students' learning progresses. Digitalization has particularly advanced the field of student assessment. Replacing pen-and-paper tests (PPT), digitally administered computer-based tests (CBT) and computer adaptive tests (CAT) have recently increased in popularity. For instance, large-scale assessments (LSA) like the *Progress of International Reading Literacy Study* (PIRLS) are switching to CBTs (Hußmann et al., 2017), while the *Programme for International Student Assessment* (PISA) is adopting adaptive tests (Yamamoto et al, 2019). CATs estimate an examinee's ability throughout the test and use this estimate to administer items that provide the highest amount of information to estimate their ability, making them more efficient than fixed-item tests (FIT; e.g. Ling et al., 2017). However, doubts about the direct comparability of these formats with regard to the test experience have been raised (e.g., Chua, 2012; Colwell, 2013).

Most research on test experience compares only two of the three test formats (PPT, CBT, or CAT) simultaneously. For instance, Chua (2012) compared the test motivation experienced in PPTs and CBTs and Martin and Lazendic (2018) investigated test anxiety in CBTs and CATs. In contrast, the effects of test formats on the test experience of elementary school students taking tests of reading comprehension, which is an important predictor for educational success (e.g. Schwabe & McElvany, 2015), are fairly unexplored. This study presents a systematic experimental study investigating the effects of different test formats on test experience among elementary school students during and after a reading comprehension test.

Test Experience

When confronted with a test with a given time frame in a classroom setting, students can experience test anxiety (von der Embse, 2018) as well as test motivation (Chua, 2012; Weiss & Betz, 1973). Test anxiety refers to a set of emotional, physiological, and behavioral responses that accompany a person's concerns about possible negative consequences of failure on a test or other evaluative situation (Sieber et al., 1977). Test motivation or test-taking motivation is a specific form of achievement motivation and can be conceptualized as a situation-specific motivation to perform well in an evaluative situation (Baumert & Demmrich, 2001). During a test, both test anxiety and test motivation are tied to the specific situation (state) and are dependent on people's predispositions (traits). While states are

situation-specific and bound to a particular point in time, traits are stable attributes (Tremblay et al., 1995). When investigating the effects of test formats on states, it is important to consider the role of the associated traits (e.g. Tremblay et al., 1995; Zohar, 1998).

Test Anxiety

In the additive model of test anxiety, test anxiety is understood as consisting of state and trait test anxiety (Zohar, 1998). State test anxiety can be influenced by several factors, such as the perceived importance of the test, the examinee's preparedness or level of self-confidence. An individual's state test anxiety results from their trait test anxiety as well as situation-specific variables. Higher levels of trait test anxiety lead to higher levels of state test anxiety (Paulman & Kennelly, 1984).

Differences in Test Anxiety Between Test Media

Older studies reported that CBT can elicit test anxiety, often in conjunction with computer anxiety (e.g. Shermis & Lombard, 1998). However, the now-ubiquitous presence of computers has made computer anxiety less relevant for children today (dos Santos & de Santana, 2018). Recently, Sahlan et al. (2021) found no differences in test anxiety between PPTs and CBTs for 11- to 16-year-old high school students taking an English language test.

Differences in Test Anxiety Between Test Modes

Adaptive test administration could influence test anxiety (Colwell, 2013). In a CAT, all examinees answer approximately the same proportion of items correctly, regardless of their ability. However, examinees often expect to answer more or fewer items correctly based on their habitual expectations. This disconnect between the proportion of items answered correctly and examinees' expectations can negatively affect an examinee's test anxiety and motivation (Tonidandel et al., 2002). Ling et al. (2017) compared the effects of different configurations of CATs and FITs in a short mathematics test for grades six to eight. Participants reported lower levels of state test anxiety in the FIT condition than in the CAT. Similarly, Martin and Lazendic (2018) found that 3rd to 9th grade students experienced higher levels of state test anxiety in the CAT condition than in the CBT condition for a mathematics test. Ortner and Caspers (2011) discovered an interaction effect between trait test anxiety and administration mode. They observed that examinees with high levels of trait test anxiety performed worse in a CAT than in a FIT and concluded that adaptive tests may exhibit biases to the disadvantage of examinees with high levels of test anxiety.

In conclusion, there is reason to believe that test formats may differ in terms of how much test anxiety examinees experience. However, the effects of different test formats on experienced state anxiety among young readers are inconclusive.

Test Motivation

Similar to anxiety, state motivation is a function of trait motivation and situational characteristics (Tremblay et al., 1995). This view extends to test-taking motivation (Helm & Warwas, 2018). In the case of a reading comprehension test, state reading motivation is highly relevant. As a motivational state, it is tied to a reading task and assesses a student's willingness to engage with the test subject in the form of the reading test items. As such, a student's state reading motivation within a reading comprehension test reflects their motivation to engage with the test items and hence with the test itself (Lepper et al., 2021). State reading motivation is affected by situational characteristics and individual traits.

Differences in Test Motivation Between Test Media

There are reasons to expect that testing on a screen can improve motivation among elementary school students. Empirical studies have found that digital media can increase children's motivation to read (Picton, 2014) as well as their test-taking motivation (Chua, 2012). One probable cause is the *novelty effect* (Shin et al., 2019), which states that new experiences can generate positive attitudes simply due to their novelty. Though today's students are more accustomed to digital media, schools in Germany rarely make use of computers for testing purposes (Fraillon et al., 2020).

Differences in Test Motivation Between Test Modes

In the 1970s, Weiss and Betz (1973) argued that CATs could have beneficial effects on test motivation, as CATs challenge high-ability students more and discourage low-ability students less, thus increasing motivation. However, evidence for this view has been lacking. Both Frey et al. (2009) in a concentration test and Ortner et al. (2014) in a reasoning test found small negative effects of CATs on test motivation measured as perceived probability of success. For elementary school students, Martin and Lazendic (2018) found no significant difference between test modes on motivation in a mathematics test. Nevertheless, there are plausible arguments for effects of CATs on test motivation, although the exact nature of such effects are unclear, especially with regard to reading comprehension tests.

Current Study

Previous research on how test formats can affect the test experience of young readers taking a reading comprehension test is limited. It is important to investigate whether theoretical expectations stemming from studies with older students hold for young students who have less experience with tests. Hence, this study investigates the effects of the PPT, CBT, and CAT test formats on the test experience of fourth grade students via the following research questions:

To what extent does administering a reading comprehension test as a PPT, CBT, or CAT affect the test anxiety of fourth grade students?

To what extent does administering a reading comprehension test as a PPT, CBT, or CAT affect the reading motivation of fourth grade students?

Based on the argument by Colwell (2013) that CATs may increase test anxiety, the first hypothesis states that fourth grade students' state test anxiety will be higher in a CAT than in a PPT or CBT, with no differences between the PPT and CBT (H1). The second hypothesis states that reading motivation is lower in the PPT condition than in the CBT and CAT (H2a), based on the assumption that digital test formats increase students' motivation (Chua, 2012). Due to the *novelty effect*, it is furthermore expected that this difference is greater in the middle of the test than at its end (H2b).

Method

Participants

To investigate the hypotheses, 526 German fourth grade students from 27 classes in 12 different elementary schools in western Germany were sampled between October 2020 and December 2021. The operational sample consisted of $N = 387$ students (46.3% female) who provided parental consent. The data from the 139 students who did not have parental consent was deleted. University ethics approval was obtained (GEKTUDO_2020_26). The students' mean age was 9.53 ($SD = 0.66$) years; 13.5% were not born in Germany. Students were assigned to one of three experimental groups within their class at random ($N_{PPT} = 120$, $N_{CBT} = 135$, $N_{CAT} = 132$). There were no differences between the students in the three test formats regarding gender, $\chi^2(2) = 3.35$, $p = .187$, country of birth (native-born or immigrant), $\chi^2(2) = 4.67$, $p = .097$, test performance, $F(2, 384) = 0.05$, $p = .950$, or mid-year grade in German

language arts, $F(2, 272) = 1.73$, $p = .179$, as an indicator of students' academic performance level.

Instruments

Test Anxiety

Trait test anxiety was measured with a shortened version of the *German Test Anxiety Inventory* (TAI-G; Wacker, Jaunzeme, & Jaksztat, 2008) by Bertrams and Englert (2014). There are five items for worry and four for emotionality. State anxiety was measured with the *State-Trait Anxiety Inventory State-Kurzskala-Deutsch* (STAI-SKD), which consists of two items for worry and three for emotionality, though differential analyses of the two subscales are not recommended (Englert et al., 2011).

Reading Motivation

Trait reading motivation was measured with three statements from the reading motivation scales used in PIRLS 2016 (Hußmann et al., 2017). A fourth item ("I enjoy reading") was added to the scale. The state reading motivation scale was based on a four-item scale used in the German national supplementary test for PISA 2000 (Kunter et al., 2002) and adapted to the reading task at hand (Lepper et al., 2021).

Reading Comprehension

The *Faire und adaptive Lesekompetenzdiagnose* (FALKE) is a reading comprehension test for German third- and fourth grade students (Ludewig et al., 2021). It consists of 44 narrative and 41 expository texts with a mean length of 60.15 (SD =17.67) words. There are 69 text-based and 63 inference-based calibrated multiple-choice items. Some items share a text, though texts are not administered twice. In all conditions, 25 items were administered. For the FIT (i.e. both the PPT and CBT), items were selected based on simulations of the adaptive version of the test using the R-library catR (Magis & Raîche, 2012) in R 4.1.2 (R Core Team, 2022). Selection criteria for items were the probability of being chosen for the adaptive test and difficulty, text type, and question type. FIT item difficulties were normally distributed.

The test formats were held visually and conceptually equivalent, with the PPT printed on ISO 216 A5 sheets in landscape format, one item per page. The constraints of the CAT were also implemented for the CBT and PPT. Specifically, students were unable to return to a previous item in both the CBT and PPT versions, as this was not possible in the CAT.

To account for rapid guessing behaviour, responses in the CBT and CAT within four seconds were considered as not administered (see Wise & DeMars, 2006). Reading comprehension scores were calculated with weighted likelihood estimates (Warm, 1989) with fixed calibrated item parameters using the library catR (Magis & Raîche, 2012). The WLE reliability of the reading comprehension test was good for all test formats (PPT: $Rel._{WLE}$ = .82; CBT: $Rel._{WLE}$ = .80; CAT: $Rel._{WLE}$ = .89). More detailed descriptions of the measures can be found in the Electronic Supplementary Material, ESM 1.

Procedure

The study used both a within-subject and between-subject design. Students were tested in their classrooms by two trained test administrators. Participants were assigned to one of the three formats (PPT, CBT, and CAT) at random. Prepared at each student's desk was a tablet and a paper-based introductory questionnaire containing trait scales for anxiety and reading motivation as well as questions on socio-demographic variables. After all students had finished the introductory pre-test questionnaire, participants were informed that they would be completing a reading comprehension test. Immediately afterwards, they were asked to rate their pre-test state anxiety. Only afterwards did students receive their assigned test format and commence with the reading comprehension test. After the 12th item (midway), as well as after the last item (post-test), the state measures for anxiety and reading motivation were administered.

Analyses

In order to investigate differences between the test formats regarding test anxiety (H1), a repeated-measures analysis of covariance (ANCOVA) with midway and post-test state anxiety as repeated measures, test format as between-subjects factor and trait anxiety and pre-test state anxiety as covariates was calculated. A repeated-measures ANCOVA with state reading motivation as within-subject factor, test format as between-subjects factor, and trait reading motivation as covariate was used to test H2a and H2b. Analyses were conducted with IBM SPSS 28.

Due to low rates of missing values, listwise deletion was used. Missing values were lowest for the midway state anxiety measure (2.3%) and highest for the post-test reading motivation scale (6.5%).

Results

Manipulation Check

In order to check whether the CAT differed substantially from the FITs in the intended ways, a manipulation check was performed by looking at the mean test difficulty and average percentage of correct answers in the FITs and CAT. Difficulty should not differ between the test formats, while the variance in the percentage of correct answers should be lower in the CAT. There was no statistically significant difference in difficulty between the FITs (M = 0.96, SD = 1.25) and CAT (M = 1.16, SD = 0.85), $t(385) = 1.67$, $p = .095$. The percentage of correctly answered items between the FITs (M = 64.5 %, SD = 2.28) and the CAT (M = 60.6 %, SD = 1.96) also did not differ, $t(303) = 1.76$, $p = .079$. However, the variance in the percentage of correct answers was lower in the CAT than in the FITs, $F(1, 383) = 6.19$, $p = .013$, indicating that the CAT performed as expected.

Test Experience

Descriptive statistics for the different measures of test experience are displayed in Table 1. The scales were generally reliable (α_{min} = .67. α_{max} = .91). Pre-test anxiety was descriptively highest in the adaptive condition, but the differences were not statistically significant, $F(2, 369) = 2.78$, $p = .063$. There was no difference between the test formats regarding trait test anxiety, $F(372, 2) = 0.42$, $p = .651$, or trait reading motivation, $F(372, 2) = 0.69$, $p = .650$.

Table 1

Means (M), standard deviations (SD) and reliability coefficients (α) for the test experience measures of trait test anxiety, state test anxiety, trait reading motivation, and state reading motivation by test format and measurement point

Test Format	Measurement point	Test anxiety			Reading motivation		
		M	*SD*	α	*M*	*SD*	α
PPT	Trait	2.26	0.66	.83	3.27	0.87	.88
	State pre	1.96	0.68	.69			
	State midway	1.83	0.69	.76	2.66	0.87	.88
	State post	1.90	0.84	.85	2.71	0.96	.91
CBT	Trait	2.25	0.65	.83	3.35	0.84	.91
	State pre	1.98	0.66	.67			
	State midway	1.96	0.68	.67	2.92	0.78	.79
	State post	2.00	0.82	.78	2.79	0.93	.89
CAT	Trait	2.32	0.64	.82	3.26	0.82	.86
	State pre	2.14	0.68	.67			
	State midway	2.05	0.69	.68	3.03	0.80	.78
	State post	2.13	0.90	.82	2.90	0.96	.90
Total	Trait	2.27	0.65	.83	3.29	0.83	.89
	State pre	2.03	0.68	.68			
	State midway	1.95	0.69	.71	2.88	0.82	.82
	State post	2.02	0.86	.82	2.80	0.95	.90

Note. $N_{total} = 387$, $N_{PPT} = 120$, $N_{CBT} = 135$, $N_{CAT} = 132$; sample sizes for the descriptive statistics range from $N_{min} = 113$ to $N_{max} = 378$; state reading motivation was not measured at the pre-test measurement point.

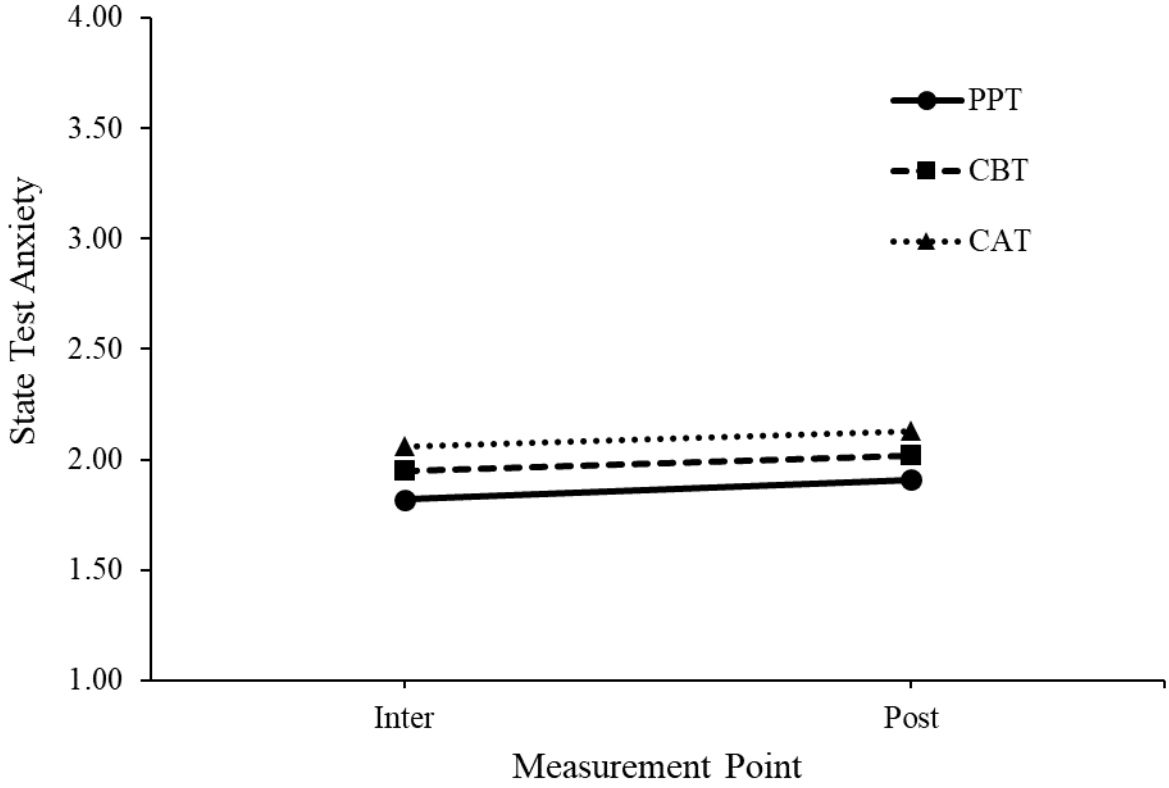
Correlations between corresponding trait and state measures were weak to moderate in the expected directions, with trait anxiety correlating weakly with the state anxiety measures (r_{min} = .356), and the state anxiety measures correlating moderately with each other (r_{min} = .539). Similarly, the correlation between trait and state reading motivation was weak (r_{min} = .329), while midway and post-test reading motivation correlated strongly with each other (r = .849). Intercorrelations among all variables can be found in the Electronic Supplementary Material, ESM 2 (Table S2).

Test Anxiety

A repeated measures ANCOVA found no effects of test format on state anxiety. There was no statistically significant difference between the test formats, $F(2, 345) = 1.88$, $p = .155$; test anxiety did not differ between measurement points, $F(1, 345) = .02$, $p = .896$; nor was there an interaction effect between state anxiety and test format, $F(2, 345) = 0.07$, $p = .935$. State anxiety was predicted by both covariates for pre-test anxiety, $F(2.345) = 73.14$, $p < .001$, and trait anxiety, $F(2, 345) = 16.94$, $p < .001$. Figure 1 shows the estimated marginal means of state anxiety over the course of the test for all three test formats. There were no statistically significant differences between the test formats regarding test anxiety, leading us to reject H1.

Figure 1

State test anxiety in the test formats over the measurement points



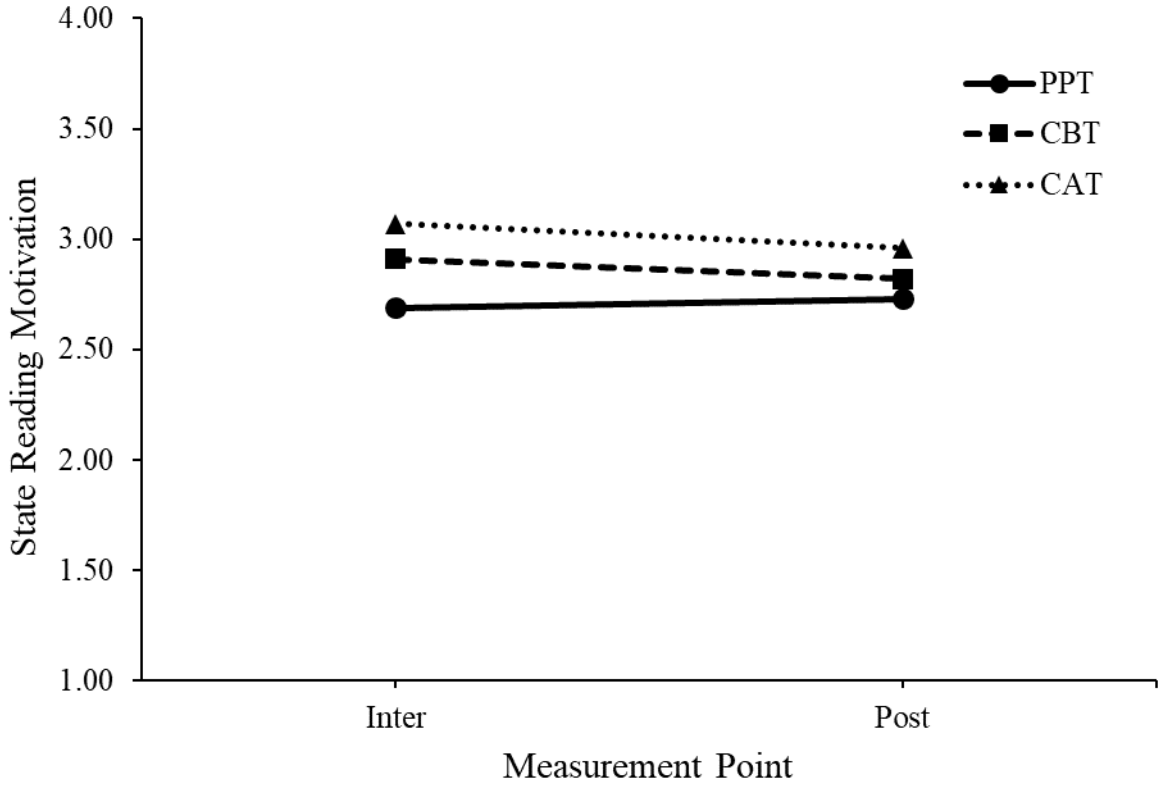
Note. This figure shows the estimated marginal means of state test anxiety for the test formats over the measurement points, controlling for pre-test anxiety and trait anxiety.

Reading Motivation

The second repeated-measures ANCOVA found that state reading motivation did not differ between the measurement points, $F(1, 339) = 2.88$, $p = .091$, but a statistically significant interaction between state reading motivation and test format indicated that state reading motivation developed differently over time between the test formats, $F(2, 339) = 3.20$, $p = .042$; $\eta_p^2 = .019$. In addition, there was a statistically significant difference between the test formats in state reading motivation, $F(2, 339) = 4.02$, $p = .019$; $\eta_p^2 = .023$. Trait reading motivation was a significant predictor for state reading motivation, $F(1, 339) = 45.27$, $p < .001$. Figure 2 shows the development of state reading motivation over the course of the test. It can be seen that state reading motivation was higher in the computer-based test formats, supporting H1a, and that motivation diminished over the course of the test in the CBT and CAT conditions, but not in the PPT condition, which is in line with H2b.

Figure 2

State reading motivation in the test formats over the measurement points



Note. This figure shows the estimated marginal means for state reading motivation for the test formats over the measurement points, controlling for trait reading motivation.
*Midway state reading motivation was significantly lower in the PPT than in the CAT.

Discussion

In this study, the effects of different test formats, namely paper-based (PPT), computer-based (CBT), and computer adaptive testing (CAT), on the test experience of 387 fourth grade students taking a reading comprehension test were investigated in a quasi-experimental within- and between-subject design. The results showed no differences in state test anxiety between the test formats. State reading motivation was initially higher when the test was administered on a screen (i.e. CBT or CAT), although the differences subsided over the course of the test.

Test Anxiety

The analyses regarding test anxiety found no statistically significant differences between the test formats or differential effects of the test formats with regards to the development of state test anxiety over the course of the test. Trait anxiety was a significant predictor for state test anxiety, which conforms to the additive model by Zohar (1998). Hypothesis 1, which stated that test anxiety would be higher in the CAT than in the FITs, had to be rejected. This result is surprising considering that previous research found that students in the adaptive testing condition experienced higher levels of test anxiety (Ling et al., 2017). It is possible that fourth grade students are less sensitive to the administration differences between a FIT and a CAT than older students. They may be less experienced with tests and have different expectations than older students, who might have stronger habitual expectations regarding their own performance and established preferences regarding test features, as described by Colwell (2013).

Reading Motivation

Reading motivation was investigated in Hypothesis H2, which assumed higher levels of reading motivation among students tested on a computer rather than on paper. The results showed statistically significantly lower levels of state reading motivation for students in the PPT at the midway point of the test, affirming H2a, although the statistically significant interaction effect indicated that this effect diminished over the course of the test, supporting H2b. This finding conforms to previous research on the motivating effects of digital media (Chua, 2012). There were no differences in test motivation between CBTs and CATs, which is in line with recent research (Martin & Lazendic, 2018). The *novelty effect* suggests that experiencing new stimuli leads to positive affect simply because the stimuli are new (Shin et al., 2019). Using computers in the classroom for testing purposes may have been a new experience for the students at first, initially increasing their reading motivation. However, over the course of the test, the students got used to the computers, causing a decline in motivation to a similar level as the students taking the PPT.

Strengths and Limitations

Data collection for this study was undertaken during the height of the COVID-19 pandemic, which forced schools to close intermittently. Thus, the tested students may have been more confident and experienced in using computers as a learning tool than fourth grade students in the past. Though this may affect the study's comparability with previous studies, it makes the results more relevant for a future in which students are more experienced with digital media for learning. The test environments were low-stakes, which might affect the results' generalizability to high-stakes test situations, but does make them relevant for large-scale assessments. Furthermore, this study did not consider differential effects of test performance on the test experience (for a brief discussion, see ESM 2). Lastly, a strength of the study was the unique experimental design comparing three test formats in parallel within a class, with within-subject measures to investigate the development of the dependant variables within individual students over time. Additionally, the manipulation check confirmed the efficacy of the administered CAT, and the within-class design allowed for comparisons even though data collection was stretched over two school years.

Conclusion

There is much we do not know about how test administration affects test experience. This study concludes that test equivalence between PPTs, CBTs and CATs in terms of test experience is achievable for young readers. CBTs and CATs do not seem to increase these students' test anxiety relative to PPTs. Instead, students are initially more motivated when being tested on a computer, though the effect wanes over time. Hence, digital media in education can yield a temporary increase in students' motivation, which can be used for instructional purposes. More efficient test formats, such as CATs, can further limit the reduction in motivation over time by allowing for shorter tests. Therefore, the results of this study should encourage more widespread use of computer-based and computer adaptive tests for reading comprehension assessment in elementary schools in low-stakes situations. Future research could look into the effects in high-stakes situations, the longevity of the motivational increase in the case of repeated computer-based test administration, and the effects of age on the test experience.

References

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462. https://doi.org/10.1007/BF03173192

Bertrams, A., & Englert, C. (2014). Test anxiety, self-control, and knowledge retrieval in secondary school students. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *46*(4), 165–170. https://doi.org/10.1026/0049-8637/a000111

Brüggemann, T. (2023). Supplement to Effects of Test Mode and Medium on Elementary School Students' Test Experience [Open Science Framework Project]. Retrieved from https://osf.io/76hc2/?view_only=baf5985b8cb94fe4b78cbce4261aee7a. https://doi.org/10.17605/OSF.IO/76HC2

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, *28*(5), 1580–1586. https://doi.org/10.1016/j.chb.2012.03.020

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies*, *1*(2), 50–60. https://doi.org/10.11114/jets.vli2.101

dos Santos, T. D., & de Santana, V. F. (2018). Computer anxiety and interaction: A systematic review. In *Proceedings of the 15th International Web for All Conference* (pp. 1-10). https://doi.org/10.1145/3192714.3192825

Englert, C., Bertrams, A., & Dickhäuser, O. (2011). Entwicklung der Fünf-Item Kurzskala STAI-SKD zur Messung von Zustandsangst [Development of the five-item short scale STAI-SKD for the assessment of state anxiety]. *Zeitschrift für Gesundheitspsychologie, 19*, 173-180. 173–180. https://doi.org/10.1026/0943-8149/a000049

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA international computer and information literacy study 2018 international report* (p. 212). Springer Nature. https://doi.org/10.1007/978-3-030-38781-5

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven

Konzentrationsleistungs-Tests [Effects of adaptive testing on test taking motivation]. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Helm, C., & Warwas, J. (2018). Psychological determinants of test motivation in low-stakes test situations: A longitudinal study of singletrait–multistate models in accounting. *Empirical Research in Vocational Education and Training*, *10*(1), 1–34. https://doi.org/10.1186/s40461-018-0074-7

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [International comparison of German elementary school students‘ reading competencies]. Münster: Waxmann.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillman, K.-J., & Weiß, M. (2002). PISA 2000: Dokumentation der Erhebungsinstrumente.: Materialien aus der Bildungsforschung Nr. 72. Max-Planck-Institut für Bildungsforschung. Lepper, C., Stang, J., & McElvany, N. (2021). Gender Differences in Text-Based Interest: Text Characteristics as Underlying Variables. *Reading Research Quarterly, 57(2)*, 537-554. https://doi.org/10.1002/rrq.420

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Ludewig, U., Trendtel, M., Schlitter, T., & McElvany, N. (2021). Adaptives Testen von Textverständnis in der Grundschule. *Diagnostica, 68*(1), 39-50.

Magis, D., & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, *48*(8). https://doi.org/10.18637/jss.v048.i08

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27-45. https://doi.org/10.1037/edu0000205

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*(3), 157-163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T.M., Weißkopf, E. & Koch, T. (2014). I Will Probably Fail. *European Journal of Psychological Assessment, 30*(1), 48-56. https://doi.org/10.1027/1015-5759/a000168

Paulman, R. G., & Kennelly, K. J. (1984). Test anxiety and ineffective test taking: Different names, same construct? *Journal of Educational Psychology*, *76*(2), 279–288. https://doi.org/10.1037/0022-0663.76.2.279

Picton, I. (2014). *The Impact of eBooks on the Reading Motivation and Reading Skills of Children and Young People: A Rapid Literature Review*. London: National Literacy Trust. Retrieved, 13 October, 2022, from https://eric.ed.gov/?id=ed560635

R Core Team (2022). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, Vienna, Austria.

Sahlan, F., Alberth, Madil, W., & Hutnisyawati (2021). The effects of modes of test administration on test anxiety and test scores: A study in an Indonesian school. *Issues in Educational Research 31*(3), 952–971.

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. Reading Research Quarterly, 50(2), 219-232.

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, *14*(1), 111–123. https://doi.org/10.1016/S0747-5632(97)00035-6

Shin, G., Feng, Y., Jarrahi, M. H., & Gafinowitz, N. (2019). Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. JAMIA open, 2(1), 62-72.

Sieber, J. E., O'Neil, J., & Tobias, S. (1977). *Anxiety, Learning, and Instruction*. Routledge. https://doi.org/10.4324/9780203056684

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320. https://doi.org/10.1037/0021-9010.87.2.320

Tremblay, P. F., Goldberg, M. P., & Gardner, R. C. (1995). Trait and state motivation and the acquisition of Hebrew vocabulary. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *27*(3), 356–370. https://doi.org/10.1037/0008-400X.27.3.356

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of affective disorders*, *227*, 483-493. https://doi.org/10.1016/j.jad.2017.11.048

Wacker, A., Jaunzeme, J., & Jaksztat, S. (2008). Eine Kurzform des Prüfungsängstlichkeitsinventars TAI-G [A short version of the test anxiety inventory TAI-G]. *Zeitschrift für Pädagogische Psychologie, 22*(1), 73-81. https://doi.org/10.1024/1010-0652.22.1.73

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/bf02294627

Weiss, D. J., & Betz, N. E. (1973). *Ability Measurement: Conventional or Adaptive?* (Research Report 73-1). Minneapolis: Department of Psychology, University of Minnesota.

Wise, S. L., & DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Yamamoto, K., Shin, H., & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. *OECD Education Working Papers*. No. 209. Paris: OECD Publishing. https://doi.org/10.1787/b9435d4b-en

Zohar, D. (1998). An additive model of test anxiety: Role of exam-specific expectations. *Journal of Educational Psychology*, *90*(2), 330–340. https://doi.org/10.1037/0022-0663.90.2.330

Effects of Test Mode and Medium on Primary School Students' Test Experience
Electronic Supplementary Material 1

Method

Test Anxiety

The State-Trait Anxiety Inventory State-Kurzskala-Deutsch (STAI-SKD) is a 5-item short scale based on the German translation of the state measure of the State-Trait Anxiety Inventory by Laux (1981). Englert et al. (2011) developed and validated a five-item state anxiety short scale for economical use based on the scale by Laux (1981). The scale encompasses the dimensions of worry and emotionality, though differential analyses of the subscales are not recommended. Items include "I am nervous" [Ich bin gerade nervös] or "I am worried" [Ich bin beunruhigt] for worry and "I am excited" [Ich bin aufgeregt] for emotionality.

Trait Test anxiety

The German Test Anxiety Inventory (TAI-G) is a nine-item measure of trait anxiety validated by Bertrams and Englert (2014). There are five items for worry and four for emotionality. Participants are asked to rate how often they experience particular emotions in test situations with statements like "I have a bad feeling during tests" [Bei Tests habe ich ein schlechtes Gefühl], "I question if my performance on a test is good enough" [Ich frage mich, ob meine Leistung im Test gut genug ist], or "During tests, I think about what would happen if I perform badly" [Bei Tests denke ich daran, was passiert, wenn ich schlecht abschneide] on a 4-point Likert scale ranging from "almost always" to "almost never".

Trait Reading Motivation

Reading motivation was measured with four statements to which participants could indicate to what extent each statement was true for them on a four-point Likert scale with the options "completely true", "somewhat true", "somewhat untrue", "completely untrue". Three statements were based on the reading motivation scale used in PIRLS 2016 (Hußmann et al., 2017). Items include "reading is fun for me" [Lesen macht mir Spaß], "reading is boring for me" [Lesen ist für mich langweilig] (reverse coded), and "I am happy when I receive a new book to read" [Ich freue mich, wenn ich ein neues Buch zum Lesen bekomme]. A fourth item, "I like reading" [Lesen gefällt mir], was added to the scale.

State reading motivation

State reading motivation was assessed during and after the reading comprehension test with a scale based on the German national supplementary test for PISA 2000 (Kunter et al., 2002). The scale was adapted to the reading task (Lepper et al., 2021) and students were asked to indicate their agreement with four items such as "The texts were interesting" [Die Texte waren interessant] or "Reading the texts is fun" [Die Texte zu lessen macht mir Spaß] on a four-point Likert scale.

Reading Comprehension

Participants were presented with 25 items in all test conditions. Text types (narrative or expository) and item types (text-based or inference-based) were counterbalanced in all test versions. In the fixed item conditions, six booklets with differently ordered items were created using a nested Latin-square design (Frey et al., 2009). Item assignment into the nested Latin squares was random. Item selection for the fixed item test was based on simulations of the adaptive version of the test using a 3-parameter logistic model (3PL) with the R-library catR (Magis & Raîche, 2012). Selection criteria for items were the probability of being chosen for the adaptive test and the text type, as well as the difficulty, text type, and question type. Item difficulties had a distribution of $N(0.95, 1.11)$. Item discriminations a and difficulties b for the FIT can be found in Table 2. The guessing parameter c was set to 0.14.

Table 2

Item parameters discrimination a and difficulty b of the items in the FIT

Item	*a*	*b*
18	1.24	0.56
23	2.33	-0.52
24	0.85	1.25
38	1.19	0.07
66	1.76	1.34
70	1.66	1.49
76	1.73	0.77
77	1.75	2.41
82	1.94	0.36
87	1.89	1.70
103	1.54	0.99
104	1.82	2.78
110	1.50	1.02
114	1.32	0.52
117	2.26	1.95
118	2.55	0.27
126	1.48	-0.19
128	0.75	0.56
136	1.74	-0.17
142	1.86	0.90
143	2.02	1.26
146	1.80	1.11
166	1.28	2.03
168	2.14	1.04
175	1.30	0.49

Note. The guessing parameter *c* of the 3PL model was set to 0.14.

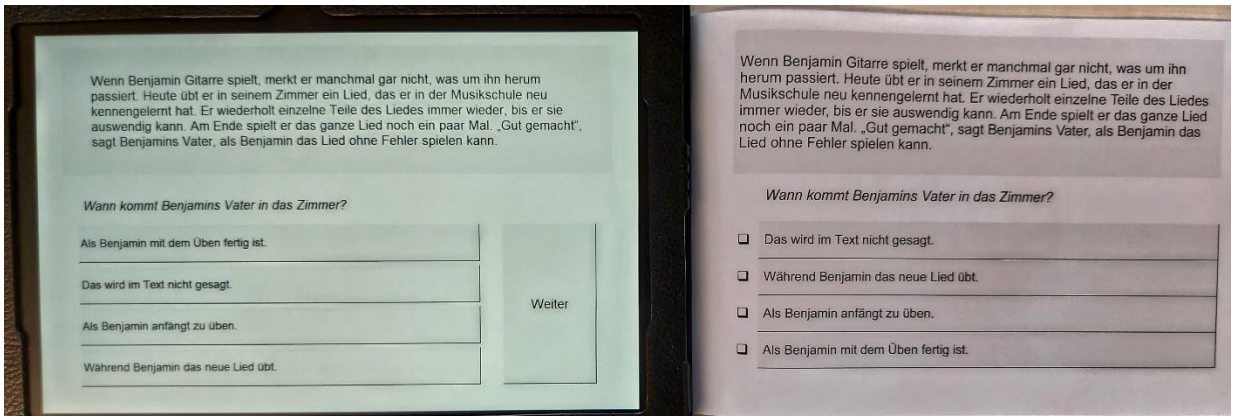
The item bank for the adaptive test based on the FALKE used all available items. The item selection criterion was the maximum Fisher information, with the examinee's ability score being estimated ad-hoc using the expected a posteriori measure (EAP; Bock & Mislevy, 1982). The prior distribution in the EAP was based on the calibration study.

The computer-based FALKE was programmed in Python 3.7 using tkinter (Lundh, 1999) for the user interface, pandas (McKinney, 2010) for data frame management, and NumPy (Harris et al., 2020) for the item selection calculations.

Test formats were held visually and conceptually equivalent. Text and item layout were similar on paper and screen. The text font used was Arial, font size for texts was 15 points (20 pixels), while items and questions were in 13.5-point (18 pixels) font. The item prompts were displayed in italics. The PPT was printed on ISO 216 A5 sheets in landscape format (10.08-inch diagonal), similar in size to the screen of the Lenovo Ideapad Miix 310-10ICR (10.1-inch diagonal), which was used for the CBT and CAT. The screen resolution was 1280 x 800 with ca. 150 PPI at 60 Hz. Figure 3 shows the display on the screen and on paper. Some constraints of the respective media limited equivalency. The display area of the PPT was adjusted to the right due to the folded area potentially affecting readability. On the screen, a button to continue to the next page was implemented to allow students to confirm their choice and limit accidental selections.

Figure 3

Item presentation on the screen (left) and on paper (right)



References

Bertrams, A., & Englert, C. (2014). Test anxiety, self-control, and knowledge retrieval in secondary school students. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *46*(4), 165–170. https://econtent.hogrefe.com/doi/pdf/10.1026/0049-8637/a000111

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Englert, C., Bertrams, A., & Dickhäuser, O. (2011). *Entwicklung der Fünf-Item-Kurzskala STAI-SKD zur Messung von Zustandsangst.* Hogrefe Verlag Göttingen. https://econtent.hogrefe.com/doi/10.1026/0943-8149/a000049

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational measurement: issues and practice*, *28*(3), 39-53.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster; New York: Waxmann. https://www.pedocs.de/frontdoor.php?source_opus=15476

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillman, K.-J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente.: Materialien aus der Bildungsforschung Nr. 72*. Max-Planck-Institut für Bildungsforschung.

Laux, L. (1981). *Das State-Trait-Angstinventar (STAI): Theoretische Grundlagen und Handanweisung*. https://opus4.kobv.de/opus4-bamberg/frontdoor/index/index/docid/31823

Lepper, C., Stang, J., & McElvany, N. (2021). Gender Differences in Text-Based Interest: Text Characteristics as Underlying Variables. *Reading Research Quarterly*. Advance online publication. https://doi.org/10.1002/rrq.420

Lundh, F. (1999). An introduction to tkinter. www.Pythonware.Com/Library/Tkinter/Introduction/Index.Html

Magis, D., & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, *48*(8). https://doi.org/10.18637/jss.v048.i08

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the Python in Science Conference, Proceedings of the 9th Python in Science Conference* (pp. 56–61). SciPy. https://doi.org/10.25080/Majora-92bf1922-00a

Electronic Supplementary Material 2

Effects of Test Mode and Medium on Primary School Students' Test Experience

Test Experience

The inter-correlations of the test experience measures as well as the reading comprehension test scores are presented in Table 3. The same scales administered at different measurement points correlated moderately to strongly with each other. Additionally, theoretically linked constructs, such as trait and state anxiety and trait and state reading motivation, displayed weak correlations. Previous studies found strong evidence for a negative relationship between test anxiety and test performance (von der Embse et al., 2018). This relationship was also found in the present sample. Furthermore, in line with Guthrie & Wigfield (2005), reading comprehension test scores were also positively correlated with trait reading motivation.

Table 3

Inter-correlations of the measures for test experience and reading comprehension

			1	2	3	4	5	6	7
1	Trait anxiety								
2	State test anxiety	Pre	**.44**						
3		Midway	**.44**	**.59**					
4		Post	**.36**	**.54**	**.73**				
5	Trait reading motivation		**-.18**	**-.11**	-.06	.037			
6	State reading motivation	Midway	.10	**.19**	**.11**	**.175**	**.35**		
7		Post	.06	**.11**	.08	**.105**	**.33**	**.85**	
8	Reading comprehension		**-.22**	**-.20**	**-.19**	**-.17**	**.33**	.06	.09

Note. Correlations in bold are statistically significant at the $p = .05$ level. Pairwise correlations range from $N_{min} = 347$ to $N_{max} = 375$.

Reading Comprehension and Test Experience

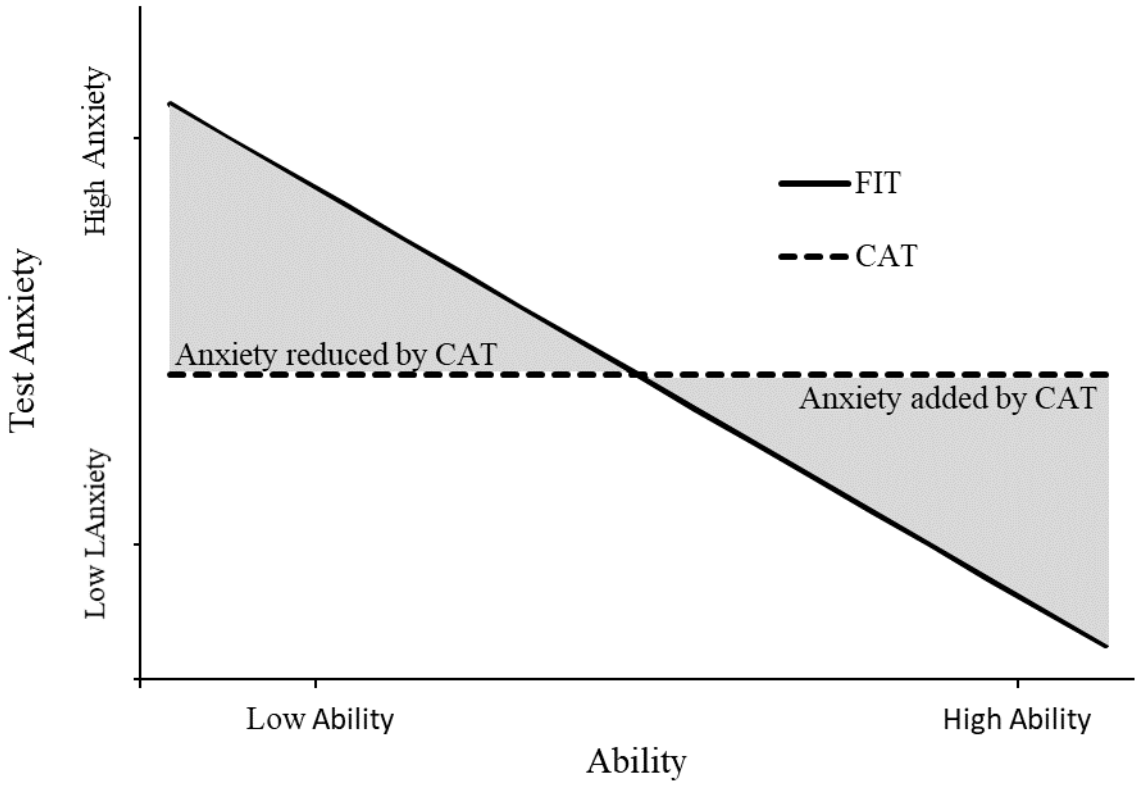
While not a focus of this study, the WLE estimates for mean test performance on the test formats in the distribution of $N(0.95, 1.11)$ were 1.28 ($SD = 1.10$) for the PPT, 1.30 ($SD = 1.14$) for the CBT, and 1.33 ($SD = 1.45$) in the CAT. Standard deviations were higher in the

CAT, because the adaptive test allowed for more extreme results among highly skilled or unskilled readers. Students in the sample were slightly more capable readers than expected from the calibration study. A one-way analysis of variance (ANOVA) found no differences in test performance between the three test formats, $F(2, 384) = 0.05$, $p = .950$.

There is an ongoing discussion on the relationship between test performance and test anxiety (von der Embse et al., 2018), especially with regards to computer adaptive testing (Colwell, 2013). Assuming an interference model of test anxiety (see Sommer & Arendasy, 2014), where high levels of test anxiety interfere with task performance, an argument can be made that high performing students would feel more challenged and thus more anxious in a CAT instead of a FIT. Conversely, low-ability students would be able to accomplish more than they expected, potentially reducing their anxiety. These effects can cancel out on the test level, as displayed in figure 4.

Figure 4

Theoretical model of the interaction between state test anxiety and test performance in a FIT and a CAT



In order to investigate this potential effect, the correlations between test performance and post-test test anxiety were compared. In the FIT (both PPT and CBT), the correlation was -.227, in the CAT -.087. Though the correlation was descriptively lower in the CAT than in the FIT, as expected by the model assumed in Figure 4, there was no statistically significant difference between the correlations, $z = 1.27$, $p = .103$. While this result suggests no differential effects of performance on test anxiety, this analysis has the strong limitation that pre-test state anxiety and trait anxiety could not be controlled for. Hence, further research in this area is required.

References

Colwell, N. M. (2013). Test Anxiety, Computer-Adaptive Testing and the Common Core. *Journal of Education and Training Studies*, *1*(2), 50-60.

Guthrie, J. T., & Wigfield, A. (2005). Roles of motivation and engagement in reading comprehension assessment. *Children's Reading Comprehension and Assessment*, 205–232.

Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence*, *42*, 115-127.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of affective disorders*, *227*, 483-493

4. Diskussion

Diese Dissertation befasst sich mit den Effekten von digitalen Medien auf die Einstellungen von Lehrkräften und Schüler*innen. Im Fokus steht zum einen die Frage, wie Lehrkräfte über den Verlauf der COVID-19-Pandemie die Effekte von digitalen Medien auf die Schüler*innen einschätzten, zum anderen, wie Schüler*innen auf digitale Medien in Testsituationen als spezifische Unterrichtssituation (Fend, 2009) kognitiv und affektiv-motivational reagierten. In den empirischen Beiträgen wurden diese Forschungsschwerpunkte untersucht.

Beitrag I befasste sich mit der Fragestellung, wie sich die Einschätzungen von Lehrkräften bezüglich der Auswirkungen von digitalen Medien auf die Schüler*innen sowie ihre Selbstwirksamkeit im Rahmen des TPACK-Modells währen den Schulschließungen aufgrund der COVID-19-Pandemie gewandelt hat. Beiträge II und III beschäftigten sich mit dem Erleben von digitalen Medien in Testsituationen der Schüler*innen. Im Beitrag II wurden die Effekte auf der kognitiven Ebene unter Berücksichtigung der *cognitive load theory* (CLT) untersucht. Schließlich lag der Fokus in Beitrag III darauf, inwiefern Grundschüler*innen in ihrer Angst und Motivation beeinflusst werden, wenn Tests statt papierbasiert computerbasiert beziehungsweise computeradaptiv eingesetzt werden.

Im Folgenden werden die zentralen Ergebnisse der Beiträge zusammengefasst (Kapitel 5.1) und anschließend diskutiert (Kapitel 5.2). Weiterhin werden die Limitationen und Stärken der Beiträge betrachtet (Kapitel 5.3). Daraufhin werden die Implikationen für Forschung und Praxis beschrieben (Kapitel 5.4) und abschließend wird ein Fazit gezogen (Kapitel 5.5).

4.1 Zentrale Ergebnisse der Einzelbeiträge

Die vorliegende Arbeit umfasst die drei empirischen Beiträge: *Unterricht zu Beginn und nach einem Jahr der Corona-Pandemie – Lehrkräftebefragungen zum Lernen mit digitalen Medien im Vergleich*, *Effects of Mode and Medium in Reading Comprehension Tests on Cognitive Load* und *Effects of Test Mode and Medium on Elementary School Students' Test Experience*. Nachfolgenden werden die jeweiligen theoretischen Hintergründe, zentralen Forschungsziele, Methoden und die Ergebnisse der drei Forschungsarbeiten zusammengefasst.

4.1.1 Beitrag I: Unterricht zu Beginn und nach einem Jahr der Corona-Pandemie – Lehrkräftebefragungen zum Lernen mit digitalen Medien im Vergleich

Theoretischer Hintergrund. Aufgrund der COVID-19-Pandemie kam es zwischen dem Frühjahr 2020 und den Osterferien 2021 zu zeitweisen Schließungen von Schulen in Deutschland (Freundl et al., 2021). Der Präsenzunterricht wurde durch Distanzunterricht ersetzt, bei dem digitale Medien eine zentrale Rolle einnahmen. Obgleich die Digitalisierung an Schulen in Deutschland von der Bildungspolitik gefördert wurde (KMK, 2016), lag Deutschland vor der COVID-19-Pandemie im internationalen Vergleich in deren Nutzungshäufigkeit im Unterricht zurück (Drossel et al., 2019; Fraillon et al., 2019). Vor diesem Hintergrund galt es zu untersuchen, wie Lehrkräfte zu Beginn und während der Schulschließungen digitale Medien und deren Einsatz im Unterricht einschätzen. Der Fokus lag hier besonders auf der Unterrichtsgestaltung mit digitalen Medien, der Selbstwirksamkeitseinschätzung im Rahmen des TPACK-Modells, sowie Einschätzungen der Effekte von digitalen Medien im Unterrichtskontext.

Forschungsziele. Es gab fünf zentrale Forschungsfragen, denen in dieser Studie nachgegangen wurde:

1. Wie waren die technischen Grundvoraussetzungen für das Lehren und Lernen im Distanzunterricht zu Beginn der Pandemie und zur Zeit der zweiten bundesweiten Schulschließungen knapp ein Jahr später?
2. Wie fand die Kommunikation zwischen Lehrenden und Lernenden in diesen beiden Phasen der Schulschließungen statt?
3. Lassen sich Unterschiede in den Einstellungen der Lehrkräfte zum Einsatz digitaler Medien im Unterricht zu den beiden Zeitpunkten ausmachen?
4. Gibt es Unterschiede in der Selbsteinschätzung der Kompetenzen der Lehrkräfte in Bezug auf den Einsatz digitaler Medien in Lernsituationen zu Beginn und nach einem Jahr der Pandemie?
5. Wie fand Lehren und Lernen digital gestützt während der beiden Phasen der Schulschließungen statt?

Stichprobe. Lehrkräfte wurden über soziale Medien rekrutiert an einer Online-Umfrage mit dem Programm LimeSurvey (Schmitz, 2023) teilzunehmen. Im Jahr 2020 nahmen 2810 Lehrkräfte (82 % weiblich) an der Umfrage Teil, im Jahr 2021 waren es 1774 Lehrkräfte (84 % weiblich). Die Lehrkräfte stammten aus allen 16 Bundesländern und unterrichteten an

primar-, sekundar- und berufsbildenden Schulen. Die Lehrkräfte waren 2020 im Durchschnitt 40.30 Jahre alt (*SD* = 9.71), 2021 waren sie im Durchschnitt 43.27 Jahre alt (*SD* = 9.67).

Methode. Mithilfe des Programmes LimeSurvey wurde eine Online-Umfrage programmiert. Sie beinhaltete demographische Angaben, eine Skala zum *technological-pedagogical and content knowledge* (TPACK; Koehler & Mishra, 2009; Schmidt et al., 2009), sowie viele weitere Items die verschiedene Themengebiete des Distanzunterrichtes abdeckten. Diese Themengebiete umfassten die technische Ausstattung der Lehrkräfte und der Schüler*innen (fünf Items), die Unterrichtsgestaltung mit digitalen Medien (sechs Items), Kommunikationswege der Lehrkräfte mit ihren Schüler*innen (sechs Items) und Einschätzungen, wie sich digitale Medien auf die Schüler*innen auswirken (fünf Items). Die Lehrkräfte konnten zu jedem der Statements ihre Zustimmung auf einer vier-Punkte Likert Skala die von „Stimme zu“ (1) über „Stimme eher zu“ (2) und „Stimme eher nicht zu“ (3) bis „Stimme nicht zu“ (4) reichte.

Um die beiden querschnittlich erhobenen Stichproben miteinander zu vergleichen, wurden die Antworten der Lehrkräfte zunächst gewichtet. Die Gewichte basierten auf dem Alter, dem Bundesland, der Schulform und dem Geschlecht der Lehrkräfte. Grundlage der Gewichtung waren die Angaben des statistischen Bundesamtes zu der demographischen Verteilung der Lehrkräfte an Schulen in Deutschland 2020. Mittelwertsunterschiede zwischen den Erhebungsjahren 2020 und 2021 wurden mit *t*-Tests für abhängige Stichproben geprüft.

Ergebnisse. Die Forschungsergebnisse zeigten leichte Fortschritte in verschiedener Hinsicht zum Einsatz von digitalen Medien im Unterricht während der COVID-19-Pandemie. Die technischen Grundvoraussetzungen besserten sich in Bezug auf die IT-Ausstattung von Lehrkräften und Schüler*innen, doch auch technische Schwierigkeiten nahmen leicht zu. Etwa 60 Prozent der Lehrkräfte gaben 2021 an, dass Schüler*innen sich technische Ausstattungen von der Schule leihen könnten, während dem 2020 nur 15 Prozent zustimmten. Die Kommunikation von Lehrkräften mit ihren Schüler*innen wandelte sich von Telefonaten, E-Mails und Briefen in 2020 zugunsten von Lernplattformen und virtuellen Treffen 2021. Dagegen änderten sich die Einstellungen der Lehrkräfte zu digitalen Medien nicht zwischen den Messzeitpunkten, jedoch gab es eine kleine aber statistisch signifikante Änderung in der Einschätzung der Lehrkräfte, ob digitale Medien die Motivation der Schüler*innen erhöht. Die Zustimmung der Lehrkräfte mit dieser Aussage sank von 94 Prozent in 2020 auf 91 Prozent in 2021. Die Selbsteinschätzung der Kompetenzen im Umgang mit digitalen Medien im Unterricht gemessen mithilfe einer TPACK-Skala (Schmidt et al., 2009) hingegen stieg

signifikant an, da Lehrkräfte eine erhöhte Kompetenz im Jahre 2021 angaben, als im Jahr 2020. Der Einsatz von digitalen Medien zur Unterrichtsgestaltung änderte sich ebenfalls, da Lehrkräfte 2021 deutlich öfter ihren Unterricht digital abhielten und Lernapps nutzten, als im Jahr 2020.

Schlussfolgerungen. Die Studie konnte zeigen, dass sich der Einsatz von digitalen Medien von Lehrkräften im Distanzunterricht zwischen 2020 und 2021 im Durchschnitt gewandelt hat. Unterricht fand 2021 deutlich öfter virtuell statt, digitale Medien wurden verstärkt zur Kommunikation mit den Schüler*innen eingesetzt und Lehrkräfte bedienten sich mehr an Ressourcen aus dem Internet, wie zum Beispiel an Apps zum Lernen oder zur Unterrichtsgestaltung, als 2020. Durch das aktive Nutzen von digitalen Medien stieg auch die Selbsteinschätzung der Lehrkräfte, digitale Medien im Unterricht kompetent einsetzen zu können. Die Einstellungen der Lehrkräfte gegenüber digitalen Medien hingegen änderten sich trotz des verstärkten Nutzens nur geringfügig.

4.1.2 Beitrag II: Effects of Mode and Medium in Reading Comprehension Tests on Cognitive Load

Theoretischer Hintergrund. Computerbasierte und computeradaptive Lesetests können sich stark von papierbasierten Tests unterscheiden (Delgado et al., 2018). Eigenschaften von Computerbildschirmen wie die Helligkeit, Auflösung oder Bildschirmwiederholungsrate sind Variablen, die sich auf die Wahrnehmung von Texten auswirken können. Adaptive Tests stellen zudem im Durchschnitt Aufgaben, die für Testpersonen anspruchsvoller sind, da Items, die deutlich über oder unter dem Fähigkeitsniveau der Testperson liegen, nicht im adaptiven Test ausgewählt werden. Diese Unterschiede zwischen papierbasierten, computerbasierten und computeradaptiven Tests können sich im Rahmen der *cognitive load theory* (CLT; (Leahy & Sweller, 2019) auf die kognitive Belastung von Schüler*innen auswirken. Spezifisch können sich die Darstellungsunterschiede zwischen papierbasierten und computerbasierten Tests auf die extrinsische kognitive Belastung (ECL) auswirken und die Unterschiede in der Itemselektion zwischen adaptiven und fixierten Tests auf die intrinsische kognitive Belastung (ICL).

Forschungsziele. In dieser Studie wurde untersucht, wie sich die Testformate PPT, CBT und CAT auf die kognitive Belastung von Schüler*innen der vierten Klassen in einem Lesekompetenztest auswirken. Aufgrund vorheriger Forschung zu Unterschieden zwischen papierbasierten und computerbasierten Tests (Furenes et al., 2021) wurde angenommen, dass die extrinsische kognitive Belastung am Papier niedriger ist als am Bildschirm (CBT und

CAT). Weiterhin wurde angenommen, dass die empfundene kognitive Belastung im CAT höher ist, da Schüler*innen im CAT im Durchschnitt mehr anspruchsvolle Items bearbeiten. Entsprechend wurde auch vermutet, dass die kognitive Belastung im CAT stärker ansteigt als in den fixierten Tests (PPT und CBT).

Stichprobe. Die Stichprobe bestand aus 212 Grundschüler*innen (49.50 % weiblich) der vierten Klasse in Nordrhein-Westfalen. Das durchschnittliche Alter lag bei 9.44 Jahren ($SD = 0.59$). Die Daten wurden zwischen Oktober und Dezember 2020 erhoben.

Methode. Die Schüler*innen wurden zufällig auf die Testformate PPT, CBT und CAT aufgeteilt. Der Lesekompetenztest FALKE wurde als Lesekompetenztest angewandt. In der Mitte und am Ende des Lesekompetenztests wurden die Schüler*innen mit einer vierstufigen Skala nach ihrer empfundenen extrinsischen (drei Items) und intrinsischen kognitiven Belastung (zwei Items) gefragt (Klepsch et al., 2017). Mehrere linear-mixed-effect Modelle wurden berechnet, um den Einfluss des Testmodus und des Testmediums auf die kognitive Belastung zu untersuchen. Kontrolliert wurde weiterhin für den Messzeitpunkt (Mitte und Ende der Erhebung), sowie die Leistung im Lesekompetenztest. Als Random Effects wurden Item- und Personeneffekte berücksichtigt. Im Baseline Modell wurde die kognitive Belastung nur vom Messzeitpunkt und der Testleistung vorhergesagt, im Modell 2 wurde das Testmedium (d.h. Papier oder Bildschirm) und die Interaktion zwischen Messzeitpunkt und Testmedium hinzugefügt. In Modell 3 wurde der Testmodus (d.h. FIT oder CAT) sowie die Interaktion von Messzeitpunkt und Testmodus hinzugefügt und in Modell 4 wurden sowohl Medium, als auch Modus und deren Interaktionen mit dem Messzeitpunkt zu dem Baseline Modell hinzugefügt.

Ergebnisse. Die Analysen zeigten einen signifikanten Effekt der Lesekompetenz des Messzeitpunkts auf die empfundene kognitive Belastung. Die Lesekompetenz war im PPT und CBT ein negativer Prädiktor für die kognitive Belastung. Die kognitive Belastung stieg über den Testverlauf signifikant an. Die theoretischen Annahmen zu den Unterschieden zwischen den Testformaten konnten nicht bestätigt werden. Es gab keinen Unterschied in der empfundenen kognitiven Belastung der Schüler*innen zwischen den Testmedien Papier und Bildschirm, oder dem Testmodus adaptiv oder fixiert. Allerdings wurde ein kleiner, aber signifikanter Interaktionseffekt des Testmodus und des Messzeitpunktes auf die kognitive Belastung gefunden, der anzeigte, dass die kognitive Belastung über den Testverlauf im CAT stärker anstieg als im FIT. Dieser Effekt wurde im Gesamtmodell jedoch nicht mehr gefunden.

Zusätzlich gab es Anzeichen dafür, dass die kognitive Belastung im CAT nicht von der allgemeinen Lesekompetenz beeinflusst wird.

Schlussfolgerungen. Die Ergebnisse wiesen nur kleine Unterschiede zwischen den Testformaten auf. Es gab statistische Anzeichen dafür, dass die kognitive Belastung in einem CAT über den Testverlauf hinweg stärker ansteigt als in FITs. Adaptive Tests, die als Stopp-Kriterium eine bestimmte Messpräzision anstreben, sind jedoch im Durchschnitt effizienter als fixierte Tests, sodass die Testlängen in einem CAT gegenüber FITs reduziert werden können. Aus diesem Grund legen die Ergebnisse dieser Studie nahe, die Testlängen bei adaptiven Tests durch ein Stopp-Kriterium und nicht durch eine Itemanzahl zu begrenzen. Ein interessanter Aspekt dieses Ergebnis ist auch, dass die Messpräzision in einem adaptiven Test, die durch jedes zusätzliche Item erreicht wird, Kosten in Form von kognitiver Belastung bei der Testperson mit sich bringt.

4.1.3 Beitrag III: Effects of Test Mode and Medium on Elementary School Students' Test Experience

Theoretischer Hintergrund. Ein Risiko von computerbasierten und computeradaptiven Tests sind die Auswirkungen auf das Testerleben der Schüler*innen. Vor allem computeradaptive Tests wurden in der Vergangenheit dafür kritisiert, dass sie sich negativ auf die Motivation und Testängstlichkeit von Testpersonen auswirken können, wenn diese nur etwa 50 Prozent der gestellten Aufgaben in einem Test korrekt beantworten können (Colwell, 2013; Ling et al., 2017). Gleichzeitig können sich computerbasierte Testformate auch positiv auf die Motivation von Schüler*innen auswirken (Chua, 2012; Picton, 2014). Dieser Motivationsanstieg beim Einsatz digitaler Medien ist jedoch oft nicht langfristig anhaltend (*novelty effect*; (Shin et al., 2019). Dabei stehen Befunde zu den Effekten von computerbasierten und computeradaptiven Tests im Vergleich zu papierbasierten Tests im Grundschulbereich noch aus.

Forschungsziele. Die Studie untersuchte, inwiefern sich die Testformate PPT, CBT und CAT in einem Lesekompetenztest auf die Testängstlichkeit und Lesemotivation auf Schüler*innen der vierten Klasse auswirken. Testängstlichkeit und Lesemotivation wurden separat untersucht. Es wurde angenommen, dass die Testängstlichkeit bei in einem CAT höher ist als im PPT und CBT (H1.1) und dass die Testängstlichkeit über den Testverlauf im CAT stärker zunimmt (H1.2). Weiterhin wurde angenommen, dass die Lesemotivation niedriger im PPT als im CBT oder CAT ist (H2.1) und dass die Lesemotivation im CBT und CAT über den Testverlauf abnimmt (H2.2).

Stichprobe. Insgesamt 387 Viertklässler*innen aus Grundschulen in Nordrhein-Westfalen nahmen an der Studie zwischen Oktober 2020 und Dezember 2021 teil. Die Schüler*innen waren durchschnittlich 9.53 Jahre alt ($SD = 0.66$) und der Anteil an weiblichen Schülerinnen lag bei 46.3 Prozent.

Methode. Die Schüler*innen wurden zufällig einem der drei Testformate (PPT, CBT oder CAT) zugewiesen und diese Gruppen unterschieden sich nicht hinsichtlich des Geschlechts, der Halbjahresnote im Fach Deutsch, oder dem Herkunftsland. Vor dem Test wurden alle Schüler*innen am Papier nach ihrer allgemeinen Testängstlichkeit und ihrer allgemeinen Lesemotivation gefragt. Weiterhin wurden demographische Variablen (Alter, Geschlecht, Herkunftsland) erhoben. Daraufhin wurde der Lesekompetenztest angekündigt. Bevor der Test bearbeitet wurde, wurden die Schüler*innen nach ihrer Zustandsangst gefragt. Daraufhin folgte der Lesekompetenztest, der in der Mitte unterbrochen wurde, um die Zustandsangst und Zustandslesemotivation zu erheben. Die Zustandsangst und Zustandslesemotivation wurden auch nach Abschluss des Tests noch einmal erfragt.

Zur Untersuchung der Forschungsfragen wurden zwei Ko-Varianzanalysen mit Messwiederholungen (RM-ANCOVA) genutzt. Dieses Verfahren wurde aufgrund niedriger fehlender Werte ausgewählt. In der ersten RM-ANCOVA war die Variable mit Messwiederholung die Zustandsangst, gemessen in der Mitte und am Ende des Tests. Ko-Variaten waren die allgemeine Testängstlichkeit und die Zustandsangst vor dem Test. Um die Forschungsfragen zur Lesemotivation zu untersuchen wurde ebenfalls eine RM-ANCOVA genutzt. Die Variable mit Messwiederholung war die Zustandslesemotivation in der Mitte sowie am Ende des Tests, kontrolliert wurde für die allgemeine Lesemotivation.

Ergebnisse. Die Analyse zur Testängstlichkeit zeigte keine statistisch signifikanten Unterschiede zwischen dem PPT, CBT oder CAT in der angegebenen Testängstlichkeit auf. Weiterhin gab es keine Veränderung der Testängstlichkeit über die Zeit. Stattdessen war die empfundene Zustandsangst für Schüler*innen in allen drei Testformaten stabil und relativ niedrig ($M_{Mitte} = 1.95$, $SD_{Mitte} = 0.69$; $M_{Ende} = 2.02$, $SD_{Ende} = 0.86$). Bei der Lesemotivation hingegen fand sich ein signifikanter Unterschied bei der Zustandslesemotivation in der Mitte des Tests zwischen dem PPT und dem CBT und CAT. Schüler*innen, die einen CBT oder CAT bearbeiteten gaben eine signifikant höhere Lesemotivation in der Mitte des Tests an, als die Schüler*innen, die den Test am Papier bearbeiteten, $F(2, 339) = 4.02$, $p = .019$, $\eta^2 = .023$. Diese erhöhte Lesemotivation am Computer war über den Testverlauf jedoch nicht stabil

höher und sank am Ende des Tests auf das Niveau der Lesemotivation der Schüler*innen, die am Papier getestet wurden, ab, $F(2, 339) = 3.20$, $p = .042$, $\eta^2 = .019$.

Schlussfolgerungen. Die Ergebnisse konnten die Hypothesen zur Testängstlichkeit nicht stützen. Schüler*innen der vierten Klasse erleben nicht mehr Testängstlichkeit in einem CAT als in einem PPT oder CBT. Weiterhin wurde gezeigt, dass Schüler*innen eine höhere Lesemotivation angeben, wenn sie an einen Lesekompetenztest an einem Computer machen, dies jedoch nicht langfristig anhält. Die Ergebnisse legen nahe, dass sich CBTs und CATs in der Grundschule eignen, da sie sich nicht negativ auf die Testängstlichkeit der Schüler*innen auswirken. Stattdessen wirken sich digitale Medien kurzfristig auf die Lesemotivation von Schüler*innen der vierten Klasse aus. Einen Erklärungsansatz bietet der *novelty effect,* der besagt, dass neue Stimuli aufgrund ihrer Neuheit motivierend wirken können. Der kurzfristige Motivationsschub könnte im Schulkontext mithilfe von digitalen Medien gezielt eingesetzt werden.

4.2 Diskussion der zentralen Ergebnisse

Diese Dissertation befasst sich mit der Fragestellung, wie sich der Einsatz von digitalen Medien auf motivationale und kognitive Aspekte des Erlebens von Lehrkräften und Schüler*innen auswirken. Beitrag I untersuchte die Ebene der Lehrkräfte im Unterricht, der im Kontext der COVID-19-Pandemie stattfand. Dabei wurde untersucht, wie sich die Einstellungen und Selbstwirksamkeitserwartungen über den Distanzunterricht hinweg entwickelten. Auf Seiten der Lehrkräfte konnte gezeigt werden, dass verstärkter Einsatz von digitalen Medien sich nicht auf die Einstellungen, wohl aber auf die Selbstwirksamkeitserwartungen in Form des TPACK (Koehler & Mishra, 2009) von Lehrkräften auswirken kann. Gleichzeitig wurde deutlich, dass Lehrkräfte zwischen 2020 und 2021 digitale Medien deutlich öfter im Unterricht einsetzten, sei es mit virtuellen Unterrichtsstunden, Lernapps oder Lernplattformen. Aktive Nutzung von digitalen Medien erhöht das TPACK (Tondeur et al., 2019), sodass sich die Entwicklung des TPACKs über die Messzeitpunkte hinweg mit bisherigen Befunden deckt. Es konnte weiterhin gezeigt werden, dass zu Beginn der Pandemie nur etwa ein Drittel der Lehrkräfte angaben, dass eine technische Ausstattung an ihrer Schule vorhanden war und dass die Ausleihe digitaler Geräte nur an wenigen Schulen möglich war. Dies deckt sich mit den Befunden von Freundl et al. (2021), die ebenfalls darauf hinweisen, dass die Schulen in Deutschland durch den Stand der Digitalisierung nicht hinreichend auf die Bedarfe der COVID-19-Pandemie und den damit

verbundenen Schulschließungen vorbereitet waren. Zusammenfassend lässt sich feststellen, dass die COVID-19 Pandemie die Lehrkräfte an Schulen in Deutschland vor neue Herausforderungen stellte, infolgedessen wurden die Kompetenzen der Lehrkräfte im Umgang mit digitalen Medien im Durchschnitt über die Pandemie hinweg durch die Notwendigkeit der Situation gestärkt.

In Beitrag II und III wurde jeweils das Testerleben von Schüler*innen untersucht. Ein positives Testerleben kann sich darauf auswirken, wie gut Testpersonen ihre Fähigkeiten unter Beweis stellen können (Martin & Lazendic, 2018). Dabei wurde den Forschungsfragen nachgegangen, wie sich papierbasierte, computerbasierte und computeradaptive Tests auf die kognitive Belastung, die Zustandsangst und die situative Lesemotivation von Schüler*innen der vierten Klasse auswirkten. Die bisherige Forschung konnte mehrere Unterschiede zwischen papierbasierten, computerbasierten und computeradaptiven Tests nachweisen und theoretisch formulieren. Papier und Bildschirme unterscheiden sich in der Darstellung (Köpper et al., 2016; Mayr et al., 2017) und fixierte und adaptive Tests unterscheiden sich in den im Test gestellten Items (Ling et al., 2017; Martin & Lazendic, 2018). Es wurde angenommen, dass die extrinsische kognitive Belastung aufgrund der Darstellungsunterschiede am Bildschirm höher ist (Delgado et al., 2018; Mayr et al., 2017; Wästlund et al., 2005). Die Richtung dieser Hypothese beruhte auf vorherigen empirischen Befunden zu Effekten digitaler Medien auf verschiedene Merkmale von Testpersonen. Frühere Studien fanden erhöhtes Stress- und Müdigkeitsempfinden, sowie geringeren Lernerfolg, wenn Texte am Bildschirm statt am Papier gelesen wurden (Noyes & Garland, 2003, 2008; Wästlund et al., 2005). Diese Erkenntnisse, wie auch die *screen inferiority* (Clinton, 2019; Delgado et al., 2018; Furenes et al., 2021; Kong et al., 2018) ließen sich potentiell durch eine erhöhte extrinsische kognitive Belastung aufgrund von Darstellungsunterschieden zwischen dem Papier und dem Bildschirm erklären. Diese Hypothese konnte in Beitrag II nicht bestätigt werden. Eine wahrscheinliche Ursache ist die Darstellungsart der Texte am Papier und am Bildschirm, die sich in der Studie sehr stark ähnelten, damit Befunde auf die Unterschiede zwischen dem Testmedium zurückgeführt werden können. Durch die Ähnlichkeit der Darstellungen wurde offenbar keine extrinsische Belastung erzeugt. Es ist möglich, dass die Weiterentwicklung der Bildschirmtechnologie seit den Studien von Wästlund et al. (2005) und Noyes und Garland (2008) dazu führte, dass sich deren Ergebnisse nicht replizieren ließen, da sich Bildschirmeffekte wie z.B. die Pixeldichte auf die Wahrnehmung der Texte auswirken kann. Mayr et al. (2017) fanden heraus, dass bei

einer Leseaufgabe an Bildschirmen mit einer Pixeldichte von 132 Pixel pro Inch (ppi) zu mehr Kopfschmerzen und Anstrengung führten, als die Aufgaben an Bildschirmen mit 264 ppi. Zusätzlich können Effekte der Darstellungen zwischen Papier und Computer deutlich reduziert werden, wenn das Design am Bildschirm an das Medium angepasst wird (Dawidowsky et al., 2021).

Auch eine *screen inferiority* konnte in den in diese Arbeit durchgeführten Untersuchungen nicht gefunden werden. Mögliche Gründe dafür sind die eben genannten Darstellungsunterschiede, die in den bisherigen Studien nicht immer berücksichtigt werden konnten. So kamen Delgado et al. (2018) zu dem Ergebnis, dass die Notwendigkeit zu Scrollen zu einer höheren Effektgrößte der *screen inferiority* beiträgt, wobei scrollen in der vorliegenden Studie aufgrund der kurzen Texte nicht möglich war. Weiterhin fand Clinton (2019) *screen inferiority* bei Sachtexten, nicht aber bei narrativen Texten. Der Einsatz von sowohl narrativen als auch Sachtexten im FALKE könnte die *screen inferiority* – neben dem Scrollen – daher unterdrückt haben.

Zudem wurde untersucht, wie kognitive Belastung und Fähigkeit in computerbasierten und adaptiven Tests interagieren. Angenommen wurde, dass Schüler*innen unabhängig von ihrer Fähigkeit im adaptiven Test ähnlich anspruchsvolle Items bearbeiten, während in einem fixierten Test die Differenz der Itemschwierigkeit I_b eines Items I und die Personenfähigkeit θ_j einer Person J bei Schüler*innen mit einer hohen beziehungsweise niedrigen Lesekompetenz stärker abweicht, was zu niedrigerer beziehungsweise höherer intrinsischer Belastung führen kann. Beitrag II konnte einen Trend feststellen, der mit dieser Annahme übereinstimmt. In den fixierten Tests war die intrinsische Belastung für Schüler*innen mit niedriger Lesekompetenz höher und für Schüler*innen mit hoher Lesekompetenz niedriger, während diese Beziehung im CAT nicht zu finden war. Dieses Ergebnis lässt sich hypothesenkonform mit der CLT erklären, da die *element interactivity* von Items mit Itemschwierigkeiten über der Fähigkeit ($\theta_j - I_b > 0$) für die Person J zu intrinsischer Belastung führt, während eine Person mit hoher Lesekompetenz weniger kognitive Ressourcen investieren muss und somit weniger Belastung verspürt (Noroozi & Karami, 2022). Im adaptiven Test ist die Differenz $\theta_j - I_b$ häufig näher an 0, sodass alle Testpersonen ein ähnliches Maß an kognitiver Belastung verspüren.

Aufgrund dieser Annahme, dass Testpersonen im adaptiven Test für sie im Durchschnitt anspruchsvollere Aufgaben erhalten als in fixierten Tests wurde zudem davon ausgegangen, dass die kognitive Belastung über den Testverlauf stärker ansteigt. Ein grundsätzlicher

Anstieg der kognitiven Belastung über den Testverlauf hinweg leitete sich aus der *working memory ressource depletion* Hypothese ab (Chen et al., 2018), nach der das Arbeitsgedächtnis limitiert ist und beim kontinuierlichen Nutzen abnimmt. In Beitrag II konnte eine Tendenz gefunden werden, nach der die kognitive Belastung im adaptiven Test stärker ansteigt als im fixierten Test, wobei das Ergebnis nicht konsistent in allen gerechneten Modellen war. Dennoch konnte mit der Anwendung der *working memory ressource depletion* in Testsituationen gezeigt werden, dass die Arbeitsgedächtniskapazität über den Testverlauf in allen Tests abnimmt.

In Beitrag III wurde vermutet, dass sich Tests am Bildschirm aufgrund ihrer Neuheit positiv auf die Lesemotivation der Schüler*innen auswirken (Shin et al., 2019). Diese Hypothese konnte bestätigt werden. Diese Annahme basierte auf dem *novelty effect* (Shin et al., 2019) und reiht sich zu den empirischen Befunden von Picton (2014) und Chua (2012) ein, nach denen sich digitale Medien positiv auf die Motivation von Schüler*innen auswirken können. Ein beitragsübergreifender Aspekt dieser Dissertation ist der langfristige Einfluss von digitalen Medien auf die Motivation der Schüler*innen. In Beitrag I wurden Lehrkräfte nach ihrer Einschätzung gefragt, inwiefern digitale Medien Schüler*innen motivieren können. Beitrag III untersuchte weiterhin die situative Lesemotivation von Grundschüler*innen über den Verlauf eines Lesekompetenztests hinweg. In beiden Fällen konnte zwischen den jeweiligen Messzeitpunkten eine kleine, aber statistisch signifikante Reduktion der Lesemotivation beziehungsweise der motivierenden Erwartung gefunden werden. Digitale Medien wurden im Unterrichtskontext an deutschen Schulen vor der Pandemie von nur etwa der Hälfte der Lehrkräfte regelmäßig eingesetzt (Drossel et al., 2019). Im Rahmen des Distanzunterrichts und der in Beitrag I beschriebenen verstärkten Nutzung von digitalen Medien für Unterrichtszwecke kamen Schüler*innen häufiger in Kontakt mit diesen im Unterrichtskontext. Somit ist es möglich, dass Schüler*innen zunächst durch die digitalen Medien motiviert wurden, da diese im Unterrichtskontext zuvor seltener genutzt wurden. Über das Schuljahr 2020/2021 hinweg gewöhnten sich mehr und mehr Schüler*innen an die Präsenz von digitalen Medien, wodurch die Einschätzung der Lehrkräfte zu den motivierenden Effekten dieser sank. Ähnlich verhält es sich in Beitrag III. Auch hier ist der Erklärungsansatz, dass Schüler*innen es gewohnt sind, Tests in Papierform zu bearbeiten. Digitale Medien sind in diesem Kontext ein neues Medium, wodurch die Schüler*innen zunächst motivierter waren. Weiterhin ist es möglich, dass Schüler*innen digitale Medien in erster Linie als Unterhaltungsmedium wahrnehmen. So geben über 80 Prozent der Eltern von

Grundschüler*innen an, dass diese Laptops für Videospiele nutzen (Diogo et al., 2018). Dementsprechend kann die Erwartungshaltung von Grundschüler*innen digitalen Medien gegenüber eher unterhaltungsbasiert sein. Über den Testverlauf hinweg wurde aber deutlich, dass der Test am Computer nicht unterhaltsamer wurde, sodass die Motivation der Schüler*innen sank.

Bezogen auf die Zustandsangst konnte in den Beiträgen kein Unterschied zwischen den Testformaten festgestellt werden. Unterschiede wurden zwischen dem Papier und dem Bildschirm durch Interaktionen mit der Computerangst vermutet (Shermis & Lombard, 1998). In der Kombination mit dem oben genannten Befund, dass digitale Medien auf Schüler*innen motivierend wirken können, ist es naheliegend, dass Grundschüler*innen an Schulen in Deutschland kaum bis keine Computerangst verspüren, sodass keine Unterschiede in der Testängstlichkeit zwischen dem Papier und dem Computer als Testmedium gefunden werden. In der Vergangenheit konnten Studien mit älteren Schüler*innen auch zeigen, dass computeradaptive Tests zu einer erhöhten Zustandsangst führen (Ling et al., 2017; Martin & Lazendic, 2018; Ortner et al., 2014). Als Ursache wurden habituelle Selbsteinschätzungen vermutet, da die Lösungsrate in einem CAT unabhängig von der Fähigkeit der Testperson ist (Colwell, 2013). Diese Ergebnisse konnten mit den Schüler*innen der vierten Klasse in Beitrag III nicht gefunden werden. Dabei ist zu beachten, dass die genannten Studien nicht für die allgemeine Testängstlichkeit und die situative Testängstlichkeit vor dem Test kontrollierten.

Zusammenfassend stellen die in dieser Dissertation vorgestellten Studien einen Beitrag zum Forschungsstand bezüglich des Erlebens von digitalen Medien im Unterricht von Lehrkräften und Schüler*innen dar. Es konnte gezeigt werden, dass digitale Testformate sich nicht negativ auf die Ängstlichkeit, Motivation oder kognitive Belastung von Schüler*innen auswirken. Im Gegenteil können computerbasierte Testformate kurzfristig die Lesemotivation von Grundschüler*innen erhöhen und computeradaptive Testformate können durch ihre Effizient potenziell die erlebte kognitive Belastung reduzieren. Vor allem computeradaptive Tests stellen somit eine Augmentation von schulischen Testsituationen dar, die mit analogen Medien nicht möglich sind. Um diese Testformate im Unterricht einsetzen zu können bedarf es Lehrkräfte, die digitale Medien kompetent einsetzen können. Die zeitweisen Schulschließungen und der Distanzunterricht aufgrund der COVID-19-Pandemie konnte die Selbstwirksamkeitserwartungen zum Umgang mit digitalen Medien im Unterricht der Lehrkräfte im Durchschnitt steigern.

4.3 Limitationen und Stärken

Die in dieser Dissertation diskutierten Beiträge haben eine Reihe von Limitationen, die es zu besprechen gilt. In Bezug auf Stichproben und Methoden ist zu Beitrag I anzumerken, dass die Daten nicht im Längsschnitt erhoben wurden, da Gelegenheitsstichproben genutzt wurden. Stattdessen wurden beide Stichproben anhand der Daten des statistischen Bundesamtes nach Alter, Geschlecht, Bundesland und Schulform gewichtet, sodass Vergleiche zwischen den Erhebungszeitpunkten dennoch möglich waren. Weiterhin wurden in Beitrag II Hypothesen basierend auf den Subdimensionen der kognitiven Belastung in Form der intrinsischen und extrinsischen Belastung mithilfe der Gesamtbelastung geprüft. Unter Berücksichtigung der CLT ist dies jedoch nicht problematisch, da sowohl theoretisch, als auch in der Praxis die Gesamtbelastung die Summe der intrinsischen und der extrinsischen Belastung ist (Ayres, 2006). Ein theoretisch angenommener Anstieg der extrinsischen Belastung bei konstant bleibender intrinsischer Belastung, wie bei einem Wechsel von papierbasierten zu computerbasierten Tests, führt somit zu einer Erhöhung der Gesamtbelastung um die Differenz der extrinsischen Belastung. Der gefundene erhöhte Anstieg der kognitiven Belastung über den Verlauf des adaptiven Tests hinweg lässt sich demnach nur theoretisch und in dem Beitrag nicht empirisch auf einen Anstieg der intrinsischen Belastung zurückführen. Weiterhin nutzten die Beiträge Selbsteinschätzungen der Teilnehmer*innen um die motivationalen und kognitiven Effekte der digitalen Medien in den Unterrichtssituationen festzustellen. In der Vergangenheit wurden Selbstberichte teilweise kritisiert (Ciuk et al., 2015; Kaplan et al., 2012) und physiologische Messinstrumente als Alternativen vorgeschlagen, die beispielsweise auf Cortisolmessungen, Blutdruck, Pupillendilatation oder Herzschlagrate basieren (Ayres et al., 2021; Roos et al., 2021). Diese Verfahren können aufschlussreich für Aspekte der untersuchten Konstrukte sein, die sich stärker physiologisch ausdrücken, wie zum Beispiel die Komponente der Erregung bei Testängstlichkeit.

Bezogen auf Beiträge II und III gibt es mehrere Limitationen zur Generalisierbarkeit, die jedoch durch das Forschungsdesign beabsichtigt waren. Ein zentraler Punkt bei den Erhebungen war es, Unterschiede zwischen den Testformaten PPT, CBT und CAT präzise auf die Wechsel des Testmodus beziehungsweise des Testmediums zurückführen zu können. Dadurch wurden formatspezifische Funktionen, wie zum Beispiel die Möglichkeit zu zuvor beantworteten Aufgaben zurückzukehren (*item review*), Scrolling oder Hypertexte in den Testformaten bewusst ausgeschlossen. Dies macht Generalisierungen der Studienergebnisse

für Testsituationen, in denen auf diese Unterschiede (bewusst) nicht geachtet wird, schwierig. Vor allem der Aspekt des Scrollens bei längeren Texten, der sich substanziell auf das Leseerlebnis auswirken kann (Delgado et al., 2018; Mangen & Kuiken, 2014; Piolat et al., 1997) wurde in dieser Studie eliminiert, sodass die Studienergebnisse vor allem im Grundschulbereich relevant sind, da ältere Schüler*innen im Schulkontext meist längere Texte lesen. Der Fokus bisheriger Studien lag meist auf älteren Schüler*innen (Ortner et al., 2014; Ortner & Caspers, 2011) oder Student*innen (Ling et al., 2017). Weiterhin bezogen sich die bisherigen Studien nicht auf die Lesekompetenz. Gerade diese Schnittmenge der Lesekompetenzuntersuchung in der Grundschule ist jedoch von besonderer Bedeutung, da die Lesekompetenz in der Grundschule ein wichtiger Prädiktor für den weiteren Schul- und Lebensverlauf ist (McElvany & Schwabe, 2019).

Ebenfalls ist die Testumgebung zu betrachten, die bei der Datenerhebung in Beiträgen II und III entstand. Da den Schüler*innen bekannt war, dass der Test unbenotet war, ihre Angaben anonymisiert wurden und ihre Testergebnisse nicht den Lehrkräften mitgeteilt wurden, lag nur ein geringer Leistungsdruck auf ihnen. Dies bedeutet auf der einen Seite, dass sich die Ergebnisse nicht direkt auf beispielweise benotete Testsituationen übertragen lassen, in denen die Schüler*innen möglicherweise stärker durch Testängstlichkeit und Zustandsangst beeinträchtigt sind. Auf der anderen Seite sind die Studienergebnisse für *large scale assessments* besonders relevant, da diese meist ähnliche Testumgebungen aufweisen, wie in den vorliegenden Beiträgen.

Die Berücksichtigung, dass formatspezifische Funktionen Einflüsse auf das Erleben von Texten haben können, ist auch eine Stärke der Beiträge II und III. Die gefundenen Effekte auf die kognitive Belastung (Beitrag II) und Lesemotivation (Beitrag III) lassen sich somit genau auf die Unterschiede zwischen dem Testmodus (Beitrag II) oder Testmedium (Beitrag III) zurückführen. Das Forschungsdesign, das Vergleiche zwischen papierbasierten, computerbasierten und computeradaptiven Tests erlaubte, umging dadurch konfundierende Effekte von Medium und Modus. Bisherige Studien hatten in der Vergangenheit oft Schwierigkeiten, die Ursachen für gefundene Unterschiede zwischen den Testformaten zu beschreiben (Ling et al., 2017). Bei der s*creen inferiority* ist beispielsweise noch unklar, inwiefern das Scrollen Leistungsunterschiede zwischen Texten am Papier und am Bildschirm beeinflusst (Delgado et al., 2018). Weiterhin wurden mithilfe des Innersubjektdesigns die Verläufe von kognitiver Belastung, Zustandsangst und situativer Lesemotivation über einen Testverlauf hinweg abgebildet. Dies erlaubt neue Einblicke in den Verlauf von motivationalen

und kognitiven Zuständen. Die Berücksichtigung der Eigenschaftsmerkmale in Beitrag III ist ebenfalls anzumerken, da bisherige Untersuchungen der Effekte der Testformate die Prädispositionen von Schüler*innen, Lesemotivation oder Testängstlichkeit zu verspüren, oft nicht berücksichtigten (Chua, 2012; Martin & Lazendic, 2018; Ortner & Caspers, 2011). Für Beitrag I liegt weiterhin die Stärke der großen Stichprobengrößen zu beiden Messzeitpunkten vor. Durch die Gewichtung waren Vergleiche zwischen den Jahren trotz des querschnittlichen Formats möglich.

4.4 Implikationen für Forschung und Praxis

Die Ergebnisse dieser Dissertation haben mehrere Implikationen. Mehrere theoretische Annahmen können durch die Beiträge bestärkt werden. So tragen die Ergebnisse aus Beitrag I, dass Lehrkräfte ihr TPACK im Verlauf der teilweisen Schulschließungen als höher einschätzen, zu den Befunden von (Tondeur et al., 2019) bei. Der Anstieg lässt sich über die aktive Nutzung der digitalen Medien von Lehrkräften im Unterrichtskontext erklären. Für den Einsatz von digitalen Medien im Unterricht sind die Fähigkeiten und Selbstwirksamkeitserwartungen zum Umgang mit digitalen Medien von Lehrkräften besonders wichtig (Petko et al., 2018). Da digitale Medien nicht inhärent zu mehr Lernerfolg führen, müssen Lehrkräfte in der Lage sein, diese zielführend einzusetzen. Die aktive Nutzung der digitalen Medien kann Lehrkräften in ihren Fähigkeiten bestärken. Beitrag I konnte zeigen, dass im Rahmen des Distanzunterrichts die aktive Nutzung von digitalen Medien im Unterricht substanziell zugenommen hat.

Die Ergebnisse zu den Effekten von digitalen Medien auf die Motivation der Schüler*innen in Beitrag I und III stärkt die Annahmen des *novelty effects* (Shin et al., 2019). Dass digitale Medien auf Schüler*innen motivierend wirken können, konnten vorherige Studien bereits zeigen (Chua, 2012; Lorenz et al., 2022; Picton, 2014). Der langfristige Verlust an Motivationspotential im Schulkontext zeigt, dass sich der *novelty effect* auch in Unterrichtssituationen finden lässt. Für die Praxis bedeutet dies, dass ein Einsatz von digitalen Medien im Unterricht die Schüler*innen zwar kurzzeitig zum Lesen motivieren kann, digitale Medien den motivierenden Effekt mit steigender Integration in den Unterricht jedoch verlieren werden. Daher ließe sich abwägen, ob digitale Medien nur spärlich im Unterricht genutzt werden sollten, um Schüler*innen gezielt zu motivieren, oder ob sie weitreichend in den Unterricht integriert werden sollen, um die Möglichkeiten auszuschöpfen. Aufgrund der in Beitrag III festgestellten geringen Effektgröße und dem rapiden Motivationsverlust

innerhalb von nur einer Unterrichtsstunde sind die langfristigen Potentiale, die digitale Medien in Form von z.B. digitalen Klassenbüchern, Lernplattformen oder Testeffizient mit sich führen vermutlich langfristig wertvoller. Zuvor kann der motivierende Effekt jedoch gezielt genutzt werden, um die Schüler*innen im Unterricht zum Lesen zu motivieren.

Die Ergebnisse aus Beitrag II sind bedeutsam für die Erweiterung *der cognitive load theory* um die Hypothese der *working memory ressource depletion* (Leahy & Sweller, 2019). Bisher wurde diese Hypothese vor allem zur Erklärung des *spacing effects* genutzt. Dieser Effekt beschreibt, dass Lernende sich besser an Lerninhalte erinnern können, wenn es eine Pause zwischen Lernsituationen gibt, als Lernende, die die Lerninhalte am Stück lernen (Delaney et al., 2018). Im Rahmen der *working memory ressource depletion* lässt sich der *spacing effekt* dadurch erklären, dass Lernende eine höhere kognitive Belastung empfinden, wenn sie Lerninhalte am Stück lernen (Chen et al., 2018), da die Ressourcen des Arbeitsgedächtnisses aufgebraucht werden. Während Pausen erholt sich diese Ressource. Der Befund aus Beitrag II, dass die kognitive Belastung über den Testverlauf ansteigt, trägt zur Evidenz für diese Hypothese bei, da sie den Anstieg der kognitiven Belastung erklären kann. Weiterhin hat dies die praktische Implikation, dass in einem adaptiven Test jedes zusätzlich gestellte Item mit Kosten in Form von kognitiver Belastung verbunden ist. In einem CAT erhöht jedes zusätzliche Item die Messpräzision (Embretson & Reise, 2009). Daher ist die Messpräzision in einem CAT mit der kognitiven Belastung der Testperson verknüpft. In der Praxis sind CATs, die eine gewisse Messpräzision anstreben, jedoch aufgrund ihrer Effizienz (Davey, 2011; Martin & Lazendic, 2018) meist deutlich kürzer als FITs, sodass der genannte Effekt nur bei CATs mit einer vorher festgelegten Anzahl an Items ins Gewicht fallen sollte. Damit die kognitive Belastung in einem adaptiven Test nicht höher wird als die in einem fixierten Test, sollten adaptive Test daher als Stopp-Kriterium die Messpräzision, statt einer festgelegten Anzahl an Items nutzen. Da die kognitive Belastung grundsätzlich jedoch in adaptiven und fixierten Tests über den Testverlauf anstieg, ist eine weitere Implikation für die Praxis, dass kurze Tests längeren vorgezogen werden sollten, um den Effekt der *working memory ressource depletion* auf die Arbeitsgedächtniskapazität der Schüler*innen zu beschränken. In diesem Sinne ist es ratsam, wenn möglich effiziente adaptive Tests anstelle von fixierten Tests zu nutzen, da adaptive Tests die Testlänge reduzieren können (Davey, 2011).

In dieser Hinsicht ist auch zu berücksichtigen, dass bisherige Befunde zu negativen Effekten des Lesens am Bildschirm auf die Lesekompetenz (Delgado et al., 2018), den

erlebten Stress und die Müdigkeit (Wästlund et al., 2005), sowie die Lernprozesse (Noyes & Garland, 2003, 2008) nicht repliziert werden konnten. Da die Darstellung der Aufgaben in Beiträgen II und III am Computer möglichst ähnlich zum Papier gehalten wurde, lassen sich die beschriebenen Effekte durch das Anpassen der Aufgaben an den Bildschirm in der Grundschule möglicherweise reduzieren (Dawidowsky et al., 2021). Eine Implikation für die Praxis ist daher, dass Apps und Programme, mit denen Schüler*innen Lesen sollen, an die Gegebenheiten des Darstellungsmediums angepasst werden sollten. Konkret könnte dies bedeuten, längere Texte am Bildschirm seitenweise darzustellen, anstatt mit einer Scroll-Funktion (Piolat et al., 1997), Hypertexte gezielt einzusetzen (DeStefano & LeFevre, 2007) oder Designprinzipien der CLT auch in digitalen Tests anzuwenden (Wong et al., 2012).

4.5 Fazit

Digitale Medien bieten vielversprechende Möglichkeiten den Schulunterricht zu unterstützen (Jude et al., 2020). Dabei ist zu beachten, dass digitale Medien im Sinne des SAMR-Modells bisherige Unterrichtsmethoden nicht nur substituieren, sondern mit formatspezifischen Funktionen erweitern können (Puentedura, 2006).

In dieser Dissertation wurden die Risiken einer Möglichkeit der Augmentierung von Leistungssituationen mithilfe von digitalen Medien in Form von computerbasierten und computeradaptiven Tests im Grundschulbereich untersucht. Computerbasierte Tests können Lehrkräfte durch die Automatisierung und Standardisierung von Testsituationen unterstützen (Alruwais et al., 2018) und computeradaptive Tests augmentieren den Test durch effiziente und individualisierte Aufgabenstellungen (Davey, 2011). Risiken dieser Testformate, die in vorherigen Studien bei älteren Schüler*innen und Testpersonen gefunden wurden, wie erhöhter Stress oder erhöhte Testängstlichkeit (Ling et al., 2017; Wästlund et al., 2005), konnten in den Beiträgen dieser Dissertation in der Grundschule nicht repliziert werden. Die Forschungsergebnisse weisen auf eine Empfehlung hin, digitale Testformate im Grundschulbereich im Schulkontext verstärkt zur Diagnostik und Kompetenzerfassung einzusetzen. Voraussetzung dafür sind – neben einer digitalen Infrastruktur an den Schulen – Lehrkräfte, die die nötigen Kompetenzen besitzen, um die digitalen Tests einzusetzen (Joo et al., 2018; Mishra & Koehler, 2006). Dies bedeutet, dass sie bestehende digitale Tests (wie z.B. CATse) identifizieren und nutzen können. Die Dissertation konnte zeigen, dass diese Kompetenzen über den Verlauf der COVID-19-Pandemie und den damit verbundenen zeitweisen Schulschließungen angestiegen sind. Da mit einem höheren TPACK die Nutzungsbeabsichtigung steigt (Joo et al., 2018) und durch das aktive Nutzen das TPACK gefördert werden kann (siehe auch Tondeur et al., 2019), ist es naheliegend, wenn Lehrkräfte auch nach der COVID-19-Pandemie beachsichtigen werden, digitale Medien im Unterricht einzusetzen. Vor diesem Hintergrund ist es für eine weitere Nutzung der Potenziale von computerbasierten und vor allem von computeradaptiven Tests wichtig, diese für den Grundschulunterricht zu erstellen, an Lehrkräfte zu disseminieren und ihnen den Umgang mit digitalen Testformaten näher zu bringen.

Literaturverzeichnis II

Alruwais, N., Wills, G. & Wald, M. (2018). Advantages and Challenges of Using e-Assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37. https://doi.org/10.18178/ijiet.2018.8.1.1008

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400. https://doi.org/10.1016/j.learninstruc.2006.09.001

Ayres, P., Lee, J. Y., Paas, F. & van Merriënboer, J. J. G. (2021). The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. *Frontiers in Psychology*, *12*, 702538. https://doi.org/10.3389/fpsyg.2021.702538

Chen, O., Castro-Alonso, J. C., Paas, F. & Sweller, J. (2018). Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educational Psychology Review*, *30*(2), 483–501. https://doi.org/10.1007/s10648-017-9426-2

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, *28*(5), 1580–1586. https://doi.org/10.1016/j.chb.2012.03.020

Ciuk, D., Troy, A. K. & Jones, M. C. (2015). *Measuring Emotion: Self-Reports vs. Physiological Indicators*. https://doi.org/10.2139/ssrn.2595359

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies*, *1*(2), 50–60.

Davey, T. (2011). A Guide to Computer Adaptive Testing Systems. *Council of Chief State School Officers*.

Dawidowsky, K., Holz, H., Schwerter, J., Pieronczyk, I. & Meurers, D. (2021). Development and Evaluation of a Tablet-Based Reading Fluency Test for Primary School Children. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM. https://doi.org/10.1145/3447526.3472033

Delaney, P. F., Godbole, N. R., Holden, L. R. & Chang, Y. (2018). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797. https://doi.org/10.1080/09658211.2017.1408841

Delgado, P., Vargas, C., Ackerman, R. & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. https://doi.org/10.1016/j.edurev.2018.09.003

DeStefano, D. & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, *23*(3), 1616–1641.

Diogo, A. M., Silva, P. & Viana, J. (2018). Children's use of ICT, family mediation, and social inequalities. *Issues in Educational Research*, *28*(1), 61–76.

Drossel, K., Eickelmann, B., Schaumburg, H. & Labusch, A. (2019). Nutzung digitaler Medien und Prädiktoren aus der Perspektive der Lehrerinnen und Lehrer im internationalen Vergleich. In B. Eickelmann, W. Bos, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert, M. Senkbeil & J. Vahrenhold (Hrsg.), *ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking* (S. 205–240). Waxmann.

Embretson, S. E. & Reise, S. P. (2009). *Item Response Theory for Pychologists*. Psychology Press.

Fend, H. (2009). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen*. Wiesbaden: Springer.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T. & Duckworth, D. (2019). *Preparing for life in a digital world: IEA international computer and information literacy study 2018 international report: Iea international computer and information*. Springer Nature.

Freundl, V., Stiegler, C. & Zierow, L. (2021). Europas Schulen in der Corona-Pandemie – ein Ländervergleich. *ifo Schnelldienst*, *74*(12), 41–50.

Furenes, M. I., Kucirkova, N. & Bus, A. G. (2021). A Comparison of Children's Reading on Paper Versus Screen: A Meta-Analysis. *Review of educational research*, *91*(4), 483–517. https://doi.org/10.3102/0034654321998074

Joo, Y. J., Park, S. & Lim, E. (2018). Factors influencing preservice teachers' intention to use technology: TPACK, teacher self-efficacy, and technology acceptance model. *Journal of Educational Technology & Society*, *21(3)*, 48–59.

Jude, N., Ziehm, J., Goldhammer, F., Drachsler, H. & Hasselhorn, M. (2020). *Digitalisierung an Schulen–eine Bestandsaufnahme*.

Kaplan, S., Dalal, R. S. & Luchman, J. N. (2012). Measurement of Emotions. In *Research Methods in Occupational Health Psychology* (S. 85–99). Routledge. https://doi.org/10.4324/9780203095249-15

Klepsch, M., Schmitz, F. & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology*, *8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

KMK. (2016). *Bildung in der digitalen Welt: Strategie der Kultusministerkonferenz.* Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Koehler, M. & Mishra, P. (2009). What is Technological Pedagogical Content Knowledge (TPACK)? *Contemporary Issues in Technology and Teacher Education*, *9*(1), 60–70.

Kong, Y., Seo, Y. S. & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, *123*, 138–149.

Köpper, M., Mayr, S. & Buchner, A. (2016). Reading from computer screen versus reading from paper: does it still make a difference? *Ergonomics*, *59*(5), 615–632. https://doi.org/10.1080/00140139.2015.1100757

Leahy, W. & Sweller, J. (2019). Cognitive Load Theory, Resource Depletion and the Delayed Testing Effect. *Educational Psychology Review*, *31*(2), 457–478. https://doi.org/10.1007/s10648-019-09476-2

Ling, G., Attali, Y., Finn, B. & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied psychological measurement*, *41*(7), 495–511.

Lorenz, R., Brüggemann, T., Eickelmann, B. & McElvany, N. (2022). Gelingensbedingungenfür dem Einsatz digtialer Medien in Lernsituationenin der Grundschule im Bereich Lesen - Befunde einer qualitativen Befragung von Lehrpersonen. In F. Lauermann, C. Jöhren, N. McElvany, M. Becker & H. Gaspard (Hrsg.), *Jahrbuch der Schulentwicklung Band 22: Multiperspektivität von Unterrichtsprozessen* (S. 65–93). Beltz Juventa.

Mangen, A. & Kuiken, D. (2014). Lost in an iPad. *Scientific Study of Literature*, *4*(2), 150–177. https://doi.org/10.1075/ssol.4.2.02man

Martin, A. J. & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of educational psychology*, *110*(1), 27. https://psycnet.apa.org/record/2017-17470-001

Mayr, S., Köpper, M. & Buchner, A. (2017). Effects of high pixel density on reading comprehension, proofreading performance, mood state, and physical discomfort. *Displays*, *48*, 41–49. https://doi.org/10.1016/j.displa.2017.03.002

McElvany, N. & Schwabe, F. (2019). Gender gap in reading digitally? Examining the role of motivation and self-concept. *Journal for educational research online*, *11*(1), 145–165. https://doi.org/10.25656/01:16791

Mishra, P. & Koehler, M. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Noroozi, S. & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia*, *12*(1), 1–19. https://doi.org/10.1186/s40468-022-00163-8

Noyes, J. M. & Garland, K. J. (2003). VDT versus paper-based text: Reply to Mayes, Sims and Koonce. *International Journal of Industrial Ergonomics*, *31*(6), 411–423.

Noyes, J. M. & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375.

Ortner, T. M. & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*.

Ortner, T. M., Weißkopf, E. & Koch, T. (2014). I Will Probably Fail. *European Journal of Psychological Assessment*.

Petko, D., Döbeli Honegger, B. & Prasse, D. (2018). Digitale Transformation in Bildung und Schule: Facetten, Entwicklungslinien und Herausforderungen für die Lehrerinnen- und Lehrerbildung. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *36*(2), 157–174. https://doi.org/10.25656/01:17094

Picton, I. (2014). The Impact of eBooks on the Reading Motivation and Reading Skills of Children and Young People: A Rapid Literature Review. *National Literacy Trust*.

Piolat, A., Roussey, J.-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, *47*(4), 565–589. https://doi.org/10.1006/ijhc.1997.0145

Puentedura, R. (2006). Transformation, technology, and education [Blog post]. Abgerufen am 22.02.2023 von http://hippasus.com/resources/tte/.

Roos, A.-L., Goetz, T., Voracek, M., Krannich, M., Bieg, M., Jarrell, A. & Pekrun, R. (2021). Test Anxiety and Physiological Arousal: A Systematic Review and Meta-Analysis. *Educational Psychology Review*, *33*(2), 579–618. https://doi.org/10.1007/s10648-020-09543-z

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. & Shin, T. S. (2009). Technological Pedagogical Content Knowledge (TPACK). *Journal of Research on Technology in Education*, *42*(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Schmitz, C. (2023). *LimeSurvey: An Open Source survey tool.* LimeSurvey Project.

Shermis, M. D. & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, *14*(1), 111–123. https://doi.org/10.1016/S0747-5632(97)00035-6

Shin, G., Feng, Y., Jarrahi, M. H. & Gafinowitz, N. (2019). Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open*, *2*(1), 62–72. https://doi.org/10.1093/jamiaopen/ooy048

Tondeur, J., Scherer, R., Baran, E., Siddiq, F., Valtonen, T. & Sointu, E. (2019). Teacher educators as gatekeepers: Preparing the next generation of teachers for technology integration in education. *British Journal of Educational Technology*, *50*(3), 1189–1209. https://doi.org/10.1111/bjet.12748

Wästlund, E., Reinikka, H., Norlander, T. & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, *21*(2), 377–394.