

BMARKS-LIT

„Wie wirksam sind die *bettermarks*-Funktionen? Eine systemgeleitete Literaturstudie aktueller Forschungsbefunde“

**Ausführlicher Bericht zur internen Verwendung bei *bettermarks*
(Arbeitspaket 3)**

vorgelegt von

**DFKI | Deutsches Forschungsinstitut für Künstliche Intelligenz GmbH
Educational Technology Lab unter der Leitung von Prof. Dr. Niels Pinkwart**

Autor*innen des Berichts

Dr. Benjamin Paaßen (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH)
Dr. Berit Blanc (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH)
Dr. Insa Reichow (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH)

Berlin, 31.01.2023

Inhaltsverzeichnis

Zusammenfassung	3
1 Hintergrund, Inhalt und Aufbau des Berichts	4
1.1 Ausgangssituation und Zielsetzung	4
1.2 Hintergrund	4
1.3 Methodisches Vorgehen und Aufbau des Berichts	6
2 Empirische Befunde für ausgewählte Kerneigenschaften von <i>bettermarks</i>	9
2.1 Kerneigenschaft A: Vollständigkeit der Aufgaben & Inhalte	11
2.2 Kerneigenschaft B: Intelligente Interaktionswerkzeuge	13
2.3 Kerneigenschaft C: Mikro-Adaptivität	16
2.4 Kerneigenschaft D: Makro-Adaptivität	20
2.5 Rahmenbedingungen E: Einsatz im Klassenverbund und Abdeckung des Curriculums	22
3 Empfehlungen für die weitere Ausgestaltung des <i>bettermarks</i>-Systems	25
3.1 Empfehlungen zur Systemgestaltung	25
3.2 Empfehlungen zu Rahmenbedingungen	26
4 Forschungsdesiderata und -potenziale für gemeinsame Follow-up-Projekte	27
Abbildungsverzeichnis	29
Tabellenverzeichnis	29
Literaturverzeichnis	29

Zusammenfassung

Dieser für die interne Verwendung bei *bettermarks* vorgesehene Bericht untersucht die Lernwirksamkeit der Kerneigenschaften und Rahmenbedingungen von *bettermarks* durch eine Auswertung der wissenschaftlichen Literatur. Durchsucht wurden insbesondere Metastudien zur Wirksamkeit von digitalen Mathematik-Lernumgebungen (13), Wirkungsstudien zu den Systemen ausgewählter Wettbewerber (87) und Wirkungsstudien zu intelligenten Tutoring-Systemen allgemein aus einer SCOPUS-Literaturrecherche (39). Die Auswertung ergab folgende zwölf Kernergebnisse (Kapitel 2), die den vier Kerneigenschaften von *bettermarks* und einer Rahmenbedingung zugeordnet sind, nämlich: (A) Vollständigkeit der Aufgaben und Inhalte, (B) Intelligente Interaktionswerkzeuge, (C) Mikro-Adaptivität, (D) Makro-Adaptivität und der Rahmenbedingungen (E) Einsatz im Klassenverbund und Abdeckung des Curriculums. Die Ergebnisse lauten:

- A1. Eine feinere Auflösung von Lernzielen und Aufgaben führt zu höherer Lernwirksamkeit.
- A2. Eine höhere Abdeckung korreliert nicht eindeutig mit Lernwirksamkeit.
- B1. Intelligente bzw. reichhaltige Eingabewerkzeuge sind als Lernformat wirksamer als „einfache Formate“.
- B2. Unterschiedliche Darstellungs- und Interaktionsformen können bei sorgsamer Gestaltung lernwirksamer sein.
- B3. Zusatzfrage: Die Evidenz zu Gamification-Elementen ist gemischt.
- C1. Fehlerbasiertes Feedback ist eher lernwirksam als einfaches Richtig/Falsch-Feedback.
- C2. Die Möglichkeit zur Korrektur und Reflexion eigener Antworten ist lernwirksam.
- C3. Das Anzeigen von Lösungswegen ist eher lernwirksam.
- C4. Zusätzliche Ressourcen (wie vorgerechnete Beispiele, Anleitungen und Videos) sind lernwirksam.
- D1. Das Identifizieren von Wissenslücken sowie die Vergabe gezielter Übungen, um diese Wissenslücken zu schließen, hat keinen eindeutigen Effekt.
- E1. Der Einsatz von ITS mit Beteiligung einer Lehrkraft ist im Vergleich zu traditionellem, nicht-digitalen Unterricht eher lernwirksam, wenn die Lehrkraft ausreichend geschult ist.
- E2. Eine höhere Abdeckung eines Curriculums korreliert nicht eindeutig mit Lernwirksamkeit (siehe A3.) ist aber als Rahmenbedingung, um im deutschen Markt bestehen zu können, unabdingbar.

Aus den Kernergebnissen wurden folgende Empfehlungen für die zukünftige Gestaltung von *bettermarks* (1.-4.) und die Rahmenbedingungen des Einsatzes (5.-6.) abgeleitet (Kapitel 3):

1. Der Fokus des *bettermarks*-Systems sollte weiterhin auf Mikro-Adaptivität liegen. Die mikro-adaptiven Hilfestellungen lassen sich noch ausbauen.
2. Die Makro-Adaptivität sollte in den Händen der Lernenden und Lehrenden verbleiben aber durch automatisierte Werkzeuge unterstützt werden.
3. Die hohe Qualität und Granularität der Inhalte sollte bewahrt werden.
4. In Sachen intelligente Interaktionswerkzeuge besteht kein Handlungsbedarf.
5. Die Einbindung der Lehrkräfte sollte verstärkt werden.
6. Die Inhalte für den curricularen Bedarf sollten ausgebaut werden.

Im abschließenden Kapitel 4 zeigt der Bericht auf, welche offenen Forschungsfragen mit besonderem Interesse für *bettermarks* noch verbleiben, die mit gemeinsamen Studien oder Follow-up-Projekten bearbeitet werden könnten. Insbesondere empfehlen wir, Studien auch innerhalb der *bettermarks*-Plattform zu ermöglichen, um eine evidenzbasierte, kontinuierliche Weiterentwicklung zu unterstützen.

1 Hintergrund, Inhalt und Aufbau des Berichts

1.1 Ausgangssituation und Zielsetzung

Die Geschäftsführung des adaptiven Mathematik-Lernsystems *bettermarks*, welches aktuell von rund 500.000 Schüler*innen genutzt wird, äußerte verstärktes Interesse an Learning Analytics- und Educational Datamining-Verfahren im Allgemeinen und an vertieftem Wissen zu Modellen für Intelligente Tutoring-Systeme sowie an internationaler Empirie zum Einsatz von KI-Systemen in der Schulbildung. Unter der Leitung von Prof. Dr. Niels Pinkwart übernahm das Educational Technology Lab am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) die Durchführung einer systemgeleiteten Literaturstudie zur Lernwirksamkeit verschiedener Kerneigenschaften und -funktionen von *bettermarks*. Die Kerneigenschaften des *bettermarks*-Systems sollen internationalen, empirischen Forschungsergebnissen zur Wirksamkeit ähnlicher Features und KI-Komponenten gegenübergestellt werden, so dass evidenzbasierte Empfehlungen für die weitere Ausgestaltung des *bettermarks*-Systems ausgesprochen werden können.

1.2 Hintergrund

Intelligente Tutoring-Systeme: Dieser Bericht behandelt die Wirkungsforschung zu digitalen Mathematik-Lernumgebungen in der Mittelstufe, insbesondere intelligenten Tutoring-Systemen (ITS). Als digitale Mathematik-Lernumgebung definieren wir jedes digitale System, das Lernenden die Möglichkeit gibt, mathematische Kenntnisse und Fertigkeiten zu erwerben. ITS sind besondere Lernumgebungen, die Lernenden automatisiert, personalisiert und situationsangepasst Unterstützung geben sollen. Sottolare et al. (2013) definieren ein ITS über vier Systemkomponenten: 1) Ein User Interface, über das Lernende mit dem System interagieren können, 2) ein Domänenmodell, das die Lerninhalte digital abbildet, 3) ein Lernendenmodell, das für jede*n Lernenden den aktuellen Lernzustand modelliert und 4) ein pädagogisches Modell, das auf Basis des Lernendenmodells und des Domänenmodells automatisiert eine jeweils passende Hilfestellung auswählen kann. VanLehn (2006) charakterisiert ein ITS dadurch, dass es zwei Arten von Hilfestellungen geben können muss, nämlich Hilfe bei der Auswahl der nächsten Lernaktivität („outer loop“; *bettermarks* spricht von „Makro-Adaptivität“) und Hilfe bei der erfolgreichen Bewältigung der einzelnen Lernaktivitäten („inner loop“; *bettermarks* spricht von „Mikro-Adaptivität“). *Bettermarks* erfüllt beide Definitionen (auch wenn Lernendenmodell und pädagogisches Modell in *bettermarks* nur implizit sind) und ist damit als ITS zu klassifizieren.

Pädagogische Theorien: Drei pädagogische Theorien bilden die Grundlage für ITS. Die Theorie des *Mastery Learning* geht davon aus, dass beinahe alle Menschen eine Fertigkeit oder Kenntnis meistern können („achieve mastery“), wenn Fertigkeiten- bzw. Kenntnisstand regelmäßig gemessen und Lernaktivitäten auf den gemessenen Stand angepasst werden (Guskey, 2010). Dieses Vorgehen entspricht der Makro-Adaptivität bzw. dem „outer loop“ eines ITS.

Die Theorie des *Scaffolding* (zu Deutsch: Gerüst) beschreibt dagegen, wie die Hilfestellung innerhalb einer Lernaktivität („inner loop“ bzw. Mikro-Adaptivität) am besten ausgestaltet werden sollte (Hammond und Gibbons, 2019). Das Ziel von Scaffolding ist, die externe Hilfestellung über die Zeit immer weiter zu reduzieren, bis Lernende eine Aktivität eigenständig erfolgreich abschließen können. Die Theorie der *Zone of proximal development* (ZPD) verbindet Makro- und Mikro-Adaptivität. Die ZPD ist definiert als die Menge der Lernaktivitäten, die eine bestimmte Person mit externer Hilfe (also mit

Scaffolding) aber noch nicht eigenständig erfolgreich bearbeiten kann (Hammond und Gibbons, 2019). Lernaktivitäten sollten stets so ausgewählt werden, dass sie in der ZPD liegen.

Wirkungsforschung zu digitalen Lernumgebungen: Um die Lernwirksamkeit einer digitalen Lernumgebung nachzuweisen, gilt folgender Studienaufbau als Standard (siehe z.B. Kulik und Fletcher, 2016): Eine Stichprobe von Lernenden aus der Zielgruppe wird zufällig auf zwei, etwa gleich große Gruppen aufgeteilt, eine Interventionsgruppe („treatment group“), die mit der digitalen Lernumgebung umgeht und eine Kontrollgruppe („control group“), die auf herkömmliche Weise lernt. Beide Gruppen legen einen Vortest („pre test“) ab, der den Kenntnis-/Fertigkeitsstand zu Beginn der Studie misst, lernen dann für eine gewisse Zeit und legen schließlich einen Nachtest („post test“) ab, der den Kenntnis-/Fertigkeitsstand zum Abschluss der Studie misst. Die digitale Lernumgebung gilt als wirksam, wenn Lernende in der Interventionsgruppe mehr dazu gelernt haben als in der Kontrollgruppe. Als Maß für die Lernwirksamkeit werden üblicherweise Effektstärken (Cohen’s *d* oder Hedge’s *g*) angegeben. Solche Effektstärken drücken aus, wie viele Standardabweichungen der Lernzuwachs in der Interventionsgruppe über dem Lernzuwachs in der Kontrollgruppe liegt. Effektstärken von bis zu 0.2 gelten gemeinhin als sehr klein, um 0.2 als klein, um 0.5 als moderat, um 0.8 als groß, um 1.2 als sehr groß und ab 2 als gewaltig (Sawilowsky, 2009).

Herausforderungen in Metastudien: Obwohl die meisten Wirkungsstudien dem o.g. Schema folgen, gibt es auch Unterschiede in der Studiengestaltung, die den beobachteten Effekt beeinflussen können, etwa:

- Wirkungsstudien verwenden unterschiedliche Prä- und Posttests. Effekte sind typischerweise am größten, wenn die Tests genau die Kenntnisse/Fähigkeiten messen, die die digitale Lernumgebung vermittelt. Allgemeinere, standardisierte Tests führen dagegen eher zu kleineren Effekten (Kulik und Fletcher, 2016; Ran et al., 2021).
- Studien sind unterschiedlich umfangreich. Kleinere Wirkungsstudien beobachten einzelne Klassen über wenige Stunden in Laborsituationen, während Studien ganze Schulen über ein gesamtes Schuljahr oder länger in ihrem üblichen Lernumfeld begleiten. Größere Studien zeigen üblicherweise kleinere Effekte als kleinere Studien (Steenbergen-Hu und Cooper, 2013; Kulik und Fletcher, 2016; Fang et al., 2018; Hillmayr et al., 2020; Ran et al., 2021). Allerdings sind gerade größere Studien besonders aussagekräftig für die tatsächliche Wirkung in realistischen Lehr-/Lernsituationen.
- Studien unterscheiden sich auch nach der Art, wie die digitale Lernumgebung in den Unterricht eingebettet wird. Wenn die digitale Lernumgebung herkömmliches Lernen ergänzt, zeigen sich üblicherweise größere Effekte, als wenn die Lernumgebung herkömmliches Lernen ersetzt (Kulik und Fletcher, 2016; Fang et al., 2018; Sun et al., 2021). Der zweite Aufbau ist aber für *bettermarks* interessanter, weil ja gerade ein erhöhter Lerneffekt erzielt werden soll, wenn Lernende die Zeit, die sie in Hausaufgaben mit Papier und Stift gesteckt hätten, stattdessen mit *bettermarks* verbringen.

In der Summe können diese Verzerrungen zu verwirrenden Ergebnissen führen. Zum Beispiel berichten Metastudien teilweise davon, dass digitale Lernumgebungen für Mathematik weniger wirksam seien als für andere Fächer. Tatsächlich lässt sich aber diese vermeintlich kleinere Wirkung darauf zurückführen, dass in der Mathematik anspruchsvollere Studiendesigns gewählt werden (Kulik und Fletcher, 2016). In unserem weiteren Vorgehen werden diese (und weitere) möglichen Verzerrungen bei der Beurteilung der Wirkungsforschung berücksichtigt.

1.3 Methodisches Vorgehen und Aufbau des Berichts

Ziel der Studie ist die Sichtung international publizierter, empirischer Evidenz für ausgewählte Funktionen und Komponenten des *bettermarks*-Systems, die in ähnlicher Form bei anderen adaptiven Mathe-Tutoring-Systemen zu finden sind. Anhand der Gegenüberstellung der Forschungsergebnisse und Kerneigenschaften sollen 1. fundierte Empirie für die Wirksamkeit des *bettermarks*-Systems zusammengestellt und 2. evidenzbasierte Empfehlungen für die weitere Ausgestaltung des Systems abgeleitet werden. Vor dem Hintergrund dieser Ziele wurden die nachstehenden Aufgaben durchgeführt:

1. Workshop (11.11.2023) und Gespräche (17.11.2023) zur *bettermarks*-Systemanalyse und zum Entwurf eines Recherchekonzepts
2. Literaturrecherche und -analyse zu ausgewählten Systemkomponenten bzw. Kerneigenschaften von *bettermarks*
3. Verfassen eines ausführlichen Berichts zum internen Gebrauch bei *bettermarks*.

Ergebnisse des Workshops

Im Workshop und den Gesprächen wurden vier Funktionsbereiche bzw. Kerneigenschaften (A bis D) sowie Rahmenbedingungen (E) von *bettermarks* mit insgesamt zehn Forschungsfragen zur **Lernwirksamkeit** bei Schüler*innen festgelegt, zu denen in der Literaturrecherche gezielt empirische Befunde gesucht und analysiert wurden:

Forschungsfragen zu Systemeigenschaften

- A) Vollständigkeit der Aufgaben & Inhalte
 1. Ist eine feinere Granularität von Lernzielen und Aufgaben lernwirksam?
 2. Ist eine höhere Abdeckung des Curriculums lernwirksam?
- B) Intelligente Interaktionswerkzeuge
 1. Sind reichhaltige Eingabewerkzeuge (z.B. Zeichenfunktionen) lernwirksam?
 2. Sind variable Aufgabentypen (z.B. verschiedene Darstellungs- und Interaktionsformen) lernwirksam?
- C) Mikro-Adaptivität
 1. Ist fehlerbasiertes Feedback (über reines richtig/falsch-Feedback hinaus) lernwirksam?
 2. Ist die Möglichkeit zur Korrektur und Reflexion (z.B. nach dem ersten Fehlversuch) lernwirksam?
 3. Ist das Anzeigen von Lösungswegen (nach dem zweiten Fehlschlag) lernwirksam?
 4. Sind zusätzliche Ressourcen (wie vorgerechnete Beispiele, Glossare, et al.) lernwirksam?
- D) Makro-Adaptivität
 1. Ist das Identifizieren von Wissenslücken sowie die gezielte Vergabe von Übungen, um diese Lücken zu schließen, lernwirksam?

Forschungsfrage zu Rahmenbedingungen

- E) Einsatz im Klassenverbund
 1. Ist die Kombination von ITS mit Lehrkräften lernwirksam (im Vergleich zu Lehrkräften ohne ITS oder ITS ohne Einbindung von Lehrkräften)?

Recherchekonzept

Unsere Recherche erfolgte in Hinblick auf die Beantwortung der o.g. Forschungsfragen. Sie ist in drei Zweige mit folgendem Vorgehen gegliedert:

1) Meta-Studien zur Lernwirksamkeit intelligenter Tutoring-Systeme im Mathematik-Schulunterricht

Hier betrachteten wir insbesondere die SWK- und TUM-Studien zu intelligenten Tutoring-Systemen in Deutschland sowie die dort zitierte Literatur.

2) Wirksamkeitsstudien zu ausgewählten intelligenten Tutoring-Systemen, die den Mathematik-Schulunterricht unterstützen

Hier durchsuchten wir die Literatur Wirkungsstudien zu den folgenden Systemen: *ASSISTMENTS*, *Cognitive Tutor*, *Calcularis*, *ALEKS*, *DreamBox*, *IXL*, *Area9*, *Mathegym*, *Studyly* (Klett), *Diagnose und Fördern* (Cornelsen) und *Mathe digital* (Cornelsen). Der Suchstring enthält jeweils „<Name des Systems> AND (study OR review OR meta-analysis)“. Aus den google scholar-Ergebnissen untersuchten wir jeweils die ersten 50 Suchergebnisse.

3) SCOPUS-Literaturrecherche, die dem Schema systematischer Literaturrecherchen folgt

Im Detail folgten wir hier den Recherche- und Analyseschritte englischsprachiger Forschungsliteratur nach etablierten Schritten aus dem Methodenfeld der systematischen Reviews (Newman and Gough 2020):

- Entwicklung der Fragestellung
- Design des konzeptuellen Frameworks der Literaturrecherche
- Bestimmung von In- und Exklusionskriterien
- Entwicklung der Suchstrategie
- Screening der Studien anhand der In- und Exklusionskriterien (basierend auf abstracts)
- Kodieren relevanter Studienergebnisse (anhand der Volltexte)
- Bewertung der Studienqualität
- Synthetisieren verschiedener Studienergebnisse zur Beantwortung der Forschungsfrage
- Schriftlicher Bericht der gesammelten Erkenntnisse

Dabei nahmen wir abweichend vom Vorgehen bei systematischen Reviews keine Kodierung durch mehrere Personen vor und sichteten darüber hinaus nur die 50 meistzitierten Publikationen.

Die Recherche wurde in der Datenbank SCOPUS mit Suchtermen für alle o.g. Forschungsfragen durchgeführt, um empirische Publikationen zur Wirksamkeit von unterschiedlich gestalteten ITS im Mathematikbereich von Schulen zu identifizieren (siehe Tabelle 1). Identifizierte Einträge wurden in eine Systematik (Excel) überführt, Duplikate automatisch und manuell entfernt.

Tabelle 1: Suchbegriffe

Thema	Suchbegriffe
ITS	„Intelligent tutoring system“ OR „intelligent tutor“ OR ITS OR “cognitive tutor” OR “learning environment” OR “learning platform” OR “digital learning system” OR “adaptive learning system” OR “computer-assisted learning environment” OR “adaptive learning technology”
AND	
Mathematik	Mathematic* OR math* OR algebra OR analysis OR geometry OR statistics OR arithmetic
AND	
Schule	(High OR middle OR secondary) school* OR K12*
AND	
Wirkungsforschung	Effect* OR evaluate* OR „empirical study“ OR assess*

Fehler! Ungültiger Eigenverweis auf Textmarke.2 gibt die In- und Exklusionskriterien wieder, die zunächst auf die Titel, Metadaten und Abstracts der gefundenen Veröffentlichungen angewandt wurden.

Aus den verbleibenden Veröffentlichungen wurden die 50 meistzitierten ausgewählt und der Analyse der Volltexte unterzogen, um die Publikationen den Forschungsfragen zuzuordnen.

Tabelle 2: In- und Exklusionskriterien

Kriterien	Inklusion	Exklusion
Publikationszeitraum	2000-2022	Vor 2000
Sprache	Englisch, Deutsch	Andere
Bildungslevel	Schule (Klasse 5-13)	Andere (z. B. Hochschule)
Unterrichtsfach	Mathematik	Andere (z. B. Biologie)
Publikationstyp	Peer-reviewed journal articles & conference papers	Andere (z. B. Dissertationen) Nicht-empirische Forschung,
Methodologie	Empirisch, primary research, reviews	Kommentare, Positions-papiere
Technologie	Digitale Lernplattformen, ITS	Andere
Inhalt	Adressiert wenigstens eine der o.g. Forschungsfragen	Adressiert keine Forschungsfrage

Die nachfolgende Darstellung und Ordnung der Recherche- und Analyseergebnisse erfolgt nach den fünf o.g. Kernbereichen und Rahmenbedingungen. Den Befunden werden jeweils die Kernantworten auf die Forschungsfragen vorangestellt. Alle Forschungsfragen beziehen sich auf die **Lernwirksamkeit** bei den Schüler*innen.

2 Empirische Befunde für ausgewählte Kerneigenschaften von *bettermarks*

Zweig 1: Metastudien

Zweig 2: Wettbewerber

Zweig 3: SCOPUS-Literaturrecherche

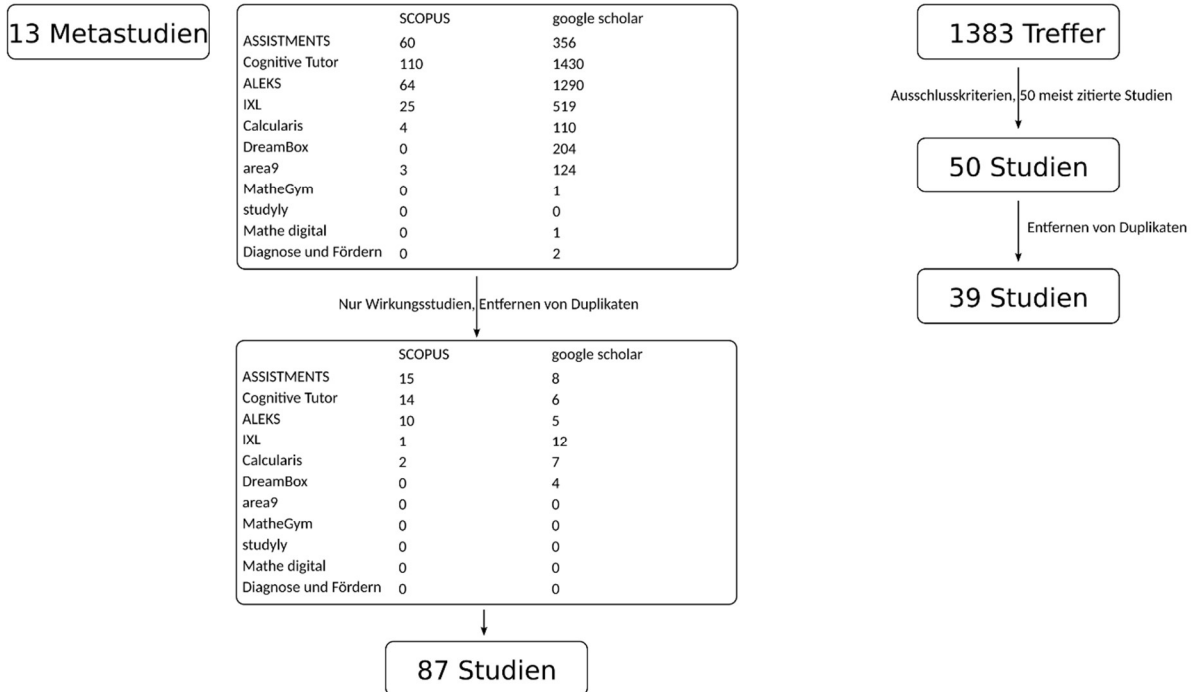


Abbildung 1: Übersicht der Fundstellen.

Abbildung 1 stellt die Fundstellen aus den einzelnen Zweigen der Recherche schematisch dar. Es wurden insgesamt 13 Metastudien aus Zweig 1 berücksichtigt, 87 Studien aus Zweig 2 und 39 Studien aus Zweig 3. Die vollständige Liste der Fundstellen ist als Anhang beigefügt.

Bemerkenswert ist, dass die Studien aus Zweig 2 sich sehr ungleich über die Wettbewerber verteilen. Zwei Drittel der Fundstellen entfallen auf nur drei Systeme, nämlich *ASSISTMENTS* (22 Fundstellen), *Cognitive Tutor* (20 Fundstellen) und *ALEKS* (15 Fundstellen). Zu *Area9*, *Mathegym*, *studyly*, *Mathe digital* und *Diagnose und Fördern* konnten wir keine einzige Wirkungsstudie finden. Daher können wir diese Systeme in den weiteren Berichten nicht berücksichtigen. Zu beachten ist ebenfalls, dass die schiere Anzahl der Befunde nicht zwingend der Wirksamkeit entspricht. Während die Wirkungsstudien zu *ASSISTMENTS*, *Cognitive Tutor* und *Calcularis* eher für deren Wirksamkeit sprechen, ist die Evidenz für die Wirksamkeit von *ALEKS*, *IXL* und *DreamBox* weniger stark (siehe Tabelle 3).

Tabelle 3: Übersicht der Wirkungsstudien zu Wettbewerbern

Belege FÜR Wirksamkeit	Belege GEGEN Wirksamkeit
ASSISTMENTS	
What Works Clearinghouse (Sahni et al., 2021); große RCT in Maine, USA zeigt positiven Effekt ($g = 0.22$; Rochelle et al., 2017); Lernwirkung auch in kleineren Studien, vgl. Prihar et al. (2022)	
Cognitive Tutor	
Besonders anspruchsvolle und groß angelegte Wirkungsstudie (Pane et al., 2016): Wirkung im zweiten Jahr der Studie; Koedinger und Aleven (2007) zitieren mehrere Metastudien aus früheren Jahren, die die Wirkung nachweisen; Barton (2016) zeigt Effekte in Subgruppen; Reed et al. (2012), Sabo et al. (2013), Aleven und Koedinger (2010) und Salden et al. (2009) zeigen Effekte für Erweiterungen	Basadre (2020) zeigt keinen Effekt; Pane (2010) zeigt keinen Effekt für den Geometrie-Tutor (Schwierigkeiten bei Lehrkräften bei der Einbindung)
Calcularis	
Positive Effekte in Rauscher et al. (2016), Kohn et al. (2020) und Gardes et al. (2022)	
ALEKS	
Positiver Effekt in den Studien von Yilmaz (2017) und Tempelaar (2012)	Kein Effekt in den Metastudie von Sun et al. (2021) und Fang et al. (2018); mehrere Einzelstudien, die ebenfalls keine Wirkung/gemischte Effekte zeigen (Fagihi et al., 2014; Fanusi, 2015; Hu et al., 2012; Kerner, 2016)
DreamBox	
Kleiner Effekt ($d = 0.14$) laut Wang und Woodworth (2011) und $d = 0.12$ laut Pellegrini et al. (2021)	Keine (überzeugenden) Effekte in den Studien von Grams (2018) und Danso (2017)
IXL	
Positive Effekte in Van Ruler (2017), Donnelly und Parmar (2022; nur bei hoher Nutzung), Arms (2019) und Palculict (2022)	Keine Effekte in den Studien von Schuetz et al. (2018), BIXLer (2019), Brown (2018) und Rholetter (2020)

Im Folgenden gehen wir nun systematisch die Kerneigenschaften des *bettermarks*-Systems sowie die Rahmenbedingungen durch und beantworten die zehn Forschungsfragen mit Hilfe der Fundstellen. Für jede Forschungsfrage berichten wir die Anzahl der Befunde, fassen die Befunde zusammen, analysieren Wettbewerber-Systeme im Hinblick auf die Forschungsfrage und schließen mit einem Fazit ab.

2.1 Kerneigenschaft A: Vollständigkeit der Aufgaben & Inhalte

Eine wesentliche Voraussetzung für das Funktionieren von *bettermarks* als ITS ist, dass Aufgaben und Inhalte zu allen Themen des Mathematik-Lehrplans und in mehreren Schwierigkeitsstufen zur Verfügung stehen (vergleiche Speroni, 2021). Die Abdeckung des Lehrplans ermöglicht es Lehrkräften, sämtliche Übungsaufgaben in *bettermarks* zu vergeben. Die Abdeckung verschiedener Schwierigkeitsstufen ermöglicht es, die Schwierigkeit der Aufgaben an das Fertigungslevel einzelner Lernender anzupassen (Makro-Adaptivität).

Im Hinblick auf die Architektur eines ITS hängt die Vollständigkeit der Aufgaben & Inhalte vor allem mit dem Domänenmodell zusammen, das die Lerninhalte und -aktivitäten der Plattform mit Kompetenzen und Wissenslücken verknüpft (Sottolare et al., 2013).

Im Kontext der Kerneigenschaft A sowie der formulierten Forschungsfragen haben wir folgende Erkenntnisse aus der Literatur ziehen können.

A1. Eine feinere Auflösung von Lernzielen und Aufgaben ist lernwirksamer.

Anzahl Befunde: 7

Kernbefunde: Aus theoretischer Perspektive begründen Ritter et al. (2007) feinschrittige Auflösungen mit der ACT-R-Theorie aus den Kognitionswissenschaften. Um prozedurales Wissen mit deklarativem Wissen zu integrieren, müssen demnach die einzelnen Aufgabenschritte Gegenstand des Lernens werden. Thurm und Graewert (2020) betrachten Mathematik-Lernsysteme aus mathematikdidaktischer Sicht und vergleichen insbesondere die Aufgabenqualität zwischen *bettermarks*, *Mathegym* und *Anton*. Als Qualitätsmerkmale betrachten sie unter anderem, dass Aufgaben mehrschrittige mathematische Modellierung oder Argumentation sowie begriffliches Denken erfordern. Diese Kriterien hängen mit feinerer Auflösung von Lernzielen und Aufgaben zusammen. Insgesamt sehen die Autor*innen **Verbesserungsbedarf beim Tiefgang der Aufgaben** in Mathematik-Lernsystemen. Im Vergleich der Systeme scheint *bettermarks* aber am ehesten positiv bewertet zu werden.

Kulik und Fletcher (2016) vergleichen Systeme, die Feedback zu Aufgaben geben („step-based feedback“), mit Systemen, die Feedback zu einzelnen Aufgabenschritten geben („substep-based feedback“) und schließen aus den Ergebnissen, dass **Feedback zu einzelnen Schritten lernwirksamer** ist. VanLehn (2011) findet 11 Studien, die step-based und substep-based tutoring direkt vergleichen und einen größeren Effekt ($d = 0.16$) für substep-based Tutoring nachweisen.

Aleven und Koedinger (2010) haben die Lernplattform *Cognitive Tutor (Geometry)* erweitert, sodass Lernende aufgefordert werden, ihre Einzelschritte zu reflektieren und zu erklären. Diese Reflexion und Erklärung führt dazu, dass weniger Aufgaben pro Zeit bearbeitet werden können, trotzdem ist die Lernwirkung im Vergleich zum regulären *Cognitive Tutor* größer, insbesondere für komplexere Aufgaben. Das spricht dafür, dass **eine feinere Auflösung von Aufgabenschritten in Kombination mit einer pädagogischen Strategie (Einladung zur Reflexion und Erklärung) lernwirksam** ist. Chang et al. (2006) untersuchen das MathCAL-Tutoring-System, das Lernende auffordert, ihre Aufgabenschritte zu berichten und weisen dessen Wirksamkeit nach (im Vergleich zu herkömmlicher Lehre).

Spitzer et al. (2021) vergleichen den Lernerfolg dreier Gruppen in den *bettermarks*-Logdaten für die Schuljahre 2019-2021: 1) Lernende, die von ihren Lehrkräften einzelne Aufgabenserien zugeteilt bekamen, 2) Lernende, die von ihren Lehrkräften umfassende Aufgabenpakete zugeteilt bekamen und 3) Lernende, die ihre Aufgaben selbst aussuchten. Insbesondere für die Zeit der Schulschließungen während der Covid-19-Pandemie im Jahr 2020, als ein Großteil der schulischen Lehre digital stattfand, zeigten sich deutliche Unterschiede. Gruppe 1 schloss mehr Aufgaben erfolgreich ab als die anderen beiden Gruppen. Das spricht dafür, dass **feiner aufgelöste Lernziele eher mit Lernerfolg korrelieren**. Der Studienaufbau erlaubt aber keinen kausalen Schluss. Es wäre z. B. möglich, dass Lehrkräfte, die kleinere Aufgabenpakete zuteilen (Gruppe 1), sowieso wirksamere Lehrkräfte sind.

***bettermarks* und Wettbewerber:** Die Systeme *Cognitive Tutor* und *Calcularis* modellieren ebenfalls feine Aufgabenschritte und modellieren Kompetenzen feingranular. Für beide Systeme liegt Wirkungsevidenz vor, was für die These spricht. *ASSISTMENTS* ist allerdings lernwirksam, obwohl keine einzelnen Aufgabenschritte modelliert werden. Umgekehrt haben *ALEKS* und *IXL* ein umfassendes und fein aufgelöstes Domänen-Modell, sind aber weniger wirksam. Aus der Analyse der Wettbewerber ergibt sich daher kein eindeutiges Bild.

Fazit: Insgesamt spricht die Wirkungsforschung eher **für** einen positiven Effekt feiner aufgelöster Aufgaben und Lernziele.

A2. Eine höhere Abdeckung eines Curriculums korreliert nicht eindeutig mit Lernwirksamkeit.

Anzahl Befunde: 4

Kernbefunde: Für den positiven Effekt einer umfassenden Abdeckung sprechen folgende Befunde: Thurm und Graewert (2020) beurteilen aus mathematik-didaktischer Sicht eine **breite Abdeckung vieler Inhalte und Aufgabentypen** als ein **Qualitätsmerkmal** von Lernsystemen. Aus Metastudien erhalten wir außerdem indirekte (korrelative) Evidenz: Kulik und Fletcher (2016) zeigen in ihrer Metastudie, dass Systeme typischerweise größere Effekte in enger gefassten Evaluationstests zeigen als in breiter gefassten, standardisierten Tests. Das deutet darauf hin, dass Systeme einen höheren Anteil des Curriculums abdecken müssen, um auch in standardisierten Tests wirksam sein zu können. Ein „What Works Clearinghouse“-Gutachten (Sahni et al., 2021) konnte nur für Systeme, die einen großen Teil des Mathematik-Curriculums abdecken, Wirksamkeit feststellen, was für eine breite Abdeckung des Curriculums spricht.

Umgekehrt spricht die Metastudie von Ran et al. (2021) eher **gegen** eine breite Abdeckung der Kompetenzen: Systeme, die mehrere Arten von Fertigkeiten adressieren, zeigen hier schwächere Lerneffekte ($d = 0.2$) als Systeme, die sich auf spezifische Fertigkeiten konzentrierten ($d = 0.58$ und höher).

***bettermarks* und Wettbewerber:** Die Analyse der Wettbewerber zeigt ebenfalls gemischte Hinweise. Für die breite Abdeckung des Curriculums spricht das Beispiel von *ASSISTMENTS*: Das System ist mit starker Wirkungsforschung validiert und deckt einen großen Teil des Mathematik-Curriculums ab. Allerdings sprechen auch mehrere Systeme eher gegen eine breite Abdeckung. *Cognitive Tutor* und *Calcularis* sind Systeme, die sich auf kleinere Teile des Curriculums (Algebra und Geometrie im Fall von *Cognitive Tutor* und Grundschul-Mathematikdidaktik für Kinder mit Dyskalkulie im Fall von *Calcularis*) konzentrieren und besonders positive Wirkungsevidenz haben. *ALEKS*, *DreamBox* und *IXL* decken eher breite Teile des Curriculums ab, haben aber wenig Wirkungsevidenz.

Fazit: Insgesamt spricht viel dafür, dass ein System mit einer breiteren Abdeckung des Curriculums grundsätzlich zu bevorzugen ist. Insbesondere erfordert der Schulkontext eine breite Abdeckung von Inhalten, damit das System auch über weite Teile eines Schuljahres einsetzbar bleibt (statt nur für einzelne Stunden und Themen). Es scheint aber **ein Spannungsverhältnis zwischen breiter Abdeckung von Inhalten und gleichzeitig hoher Aufgabenqualität** zu bestehen. Systeme, die sich eher für die Breite entscheiden (wie *ALEKS*, *DreamBox*, *IXL* und *ASSISTMENTS*) haben oft eine geringe Wirkung, während eher eng gefasste Systeme (wie *Cognitive Tutor* und *Calcularis*) in ihren jeweiligen Bereichen wirksam sind. Es gilt also, bei einer Ausdehnung der Inhalte eine weiterhin hohe Qualität der Aufgaben zu gewährleisten.

2.2 Kerneigenschaft B: Intelligente Interaktionswerkzeuge

Um mathematische Aufgaben auszuführen und zu lösen, benötigt man, neben Wissen und Fähigkeiten, unterschiedliche Werkzeuge, mit denen man dieses mathematische Wissen auch anwenden kann, wie z. B. Stift, Lineal, Zirkel, Dreieck. *bettermarks* bietet hierfür die Möglichkeit, mittels mehr als 100 verschiedenen mathematik-spezifischen Interaktionswerkzeugen von der Formeleingabe über geometrische Konstruktionen bis hin zum Zeichnen von Wahrscheinlichkeitsbäumen die Lösungen frei einzugeben (vgl. Speroni, 2021). Das System erlaubt dabei auch individuelle Lösungswege, erkennt von Fehlermuster (2.800 sind hinterlegt) und gibt unmittelbares adaptives Feedback mit passenden Hilfestellungen. Der hohe Freiheitsgrad der Werkzeuge soll die Schüler*innen zur aktiven Auseinandersetzung mit den Aufgaben motivieren, so dass sie im besten Fall ihren eigenen Lösungsweg finden. Wie lernwirksam solche multiplen Eingabewerkzeuge und Darstellungsformen sind, soll durch empirische Forschungsergebnisse belegt werden.

B1. Intelligente bzw. reichhaltige Eingabewerkzeuge sind als Lernformat wirksamer als „einfache Formate“.

Anzahl Befunde: 10

Kernbefunde: Für die größere Lernwirksamkeit reichhaltiger Eingabewerkzeuge, wie z. B. das Zeichnen von Funktionen, gegenüber „einfacher Formate“, wie z. B. Multiple Choice-Fragen, spricht eine Reihe von empirischen Studien. Gleichwohl ist festzuhalten, dass diese Studien nicht exakt die gleichen Eingabewerkzeuge untersucht, wie sie *bettermarks* anbietet, daher bieten die folgenden Befunde lediglich eine Annäherung an die Beantwortung der Forschungsfrage.

In der Metastudie zum Effekt von intelligenten Tutoring-Systemen von Ma et al. (2014) zeigte sich, dass ITS im Vergleich zu einfacher computer-basierter Instruktion eine Effektstärke von $g = 0.57$ (random effects model) erreichen. **Andere Eingabetypen als Multiple Choice** haben dabei die deutlich **höchste Lernwirksamkeit** ($g = 0.91$). Gerard et al. (2015) verglichen die Wirkung verschiedener Anleitungsmethoden und stellen fest, dass die automatisierte adaptive Anleitung am effektivsten ist, **wenn die Lernenden Ideen entwickeln und integrieren**, also z. B. Texte schreiben, Diagramme erstellen, anstatt Multiple-Choice-Aufgaben zu lösen. Auch Prihar et al. (2022) zeigen in ihrer Metastudie zu Wirkungsstudien zu *ASSISTMENTS*, zwar sehr niedrige, aber hoch signifikante Effekte, dass **offene Eingaben lernförderlicher** sind als Multiple Choice ($d = 0.085$).

Dynamische Mathe-Tools wie Geogebra ($g = 1.02$) und ITS ($g = 0.89$) sind wirksamer als einfachere Werkzeuge wie Hypermedia ($g = 0.4$) und Drill & Practice ($g = 0.58$), schließen Hillmayr et al. (2020) in

ihrer Metastudie zum Effekt von digitalen Lehr/Lernwerkzeugen im Mathe- und Wissenschaftsunterricht an Mittelschulen.

In ihrer Studie zur Wirkung von Scaffolding weisen Belland et al. (2016) darauf hin, dass die größten Effektstärken bei Aufgaben im Rahmen von **Projektlernen**, **Problemlösungen** und **Modellierung** und **Visualisierung** zu verzeichnen sind – alles Lehrbereiche bzw. Fähigkeiten, die auf reichhaltige Eingabewerkzeuge hindeuten, ohne dass deren Wirksamkeit im Einzelnen untersucht wurde. Dass webbasierte Matheinstruktionen dann vorteilhaft sind, wenn sie es dem Einzelnen ermöglichen, komplizierte Fähigkeiten wie räumliche Darstellungen durch Visualisierung zu erlernen, zeigten bereits Rafi et al. (2005, zit. nach Akin 2022).

Gegen eine positive Wirkung besonders vielfältiger Eingabemöglichkeiten sprechen zwei Befunde, die Köller et al. (2022) in ihrer umfassenden Metastudie zu Bildung mit digitalen Werkzeugen anführen: Davon ausgehend, dass das Instruktionsdesign die kognitive Belastung und lernhinderliche Prozesse minimieren sollte, finden sie reichhaltige Evidenz dafür, dass rein exploratives Lernen nicht hilft, sondern Struktur und teilweise auch direkte Instruktion durch Lehrkräfte nötig ist. Nicht nur sei für inklusive Bildung die „Passung zwischen der Funktion einer bestimmten Medieneigenschaft und dem anvisierten Kompetenzaspekt“ (ebd., S. 42) wichtig. Auch können Schüler*innen durch die vielen Handlungsoptionen oftmals kognitiv überfordert und abgelenkt sein. Diese Überforderung nach der cognitive load theory spricht gegen die Lernförderlichkeit des Einsatzes möglichst vieler unterschiedlicher Eingabewerkzeuge, ist aber vor dem Hintergrund der eingangs geschilderten Forschungslage kein eindeutiger Beleg.

***bettermarks* und Wettbewerber:** Diese Einordnung wird gemeinsam für beide Forschungsfragen am Ende von Abschnitt 2.2 vorgenommen.

Fazit: Die Forschung bietet deutliche Evidenz für die Lernwirksamkeit multipler Eingabewerkzeuge gegenüber einfachen Formaten wie Multiple Choice. Gleichwohl sind der Freiheitsgrad, die Anzahl angebotener Werkzeuge und ihre Art abzustimmen auf das jeweilige Lernziel. Es scheint empfehlenswert, sie adaptiv zum Lern- und Kompetenzstand der Schüler*innen anzubieten.

B2. Unterschiedliche Darstellungs- und Interaktionsformen können bei sorgsamer Gestaltung lernwirksamer sein.

Anzahl Befunde: 9

Kernbefunde: Zur Prüfung der Lernförderlichkeit variabler Aufgabentypen vs. einzelner Aufgabentypen pro Lernziel liegen nur wenige Befunde vor, die sich unmittelbar mit der Aufgaben- und Lernzielstruktur bei *bettermarks* vergleichen bzw. auf sie übertragen lassen. Für Thurm und Graewert (2020) sollten Aufgaben besonders die tiefergehende Auseinandersetzung fördern: „engage students at a deeper level by demanding interpretation, flexibility, the shepherding of resources, and the construction of meaning“ (Stein et al., 1996, zitiert nach Thurm und Graewert, 2020, S. 23). Eine Vielfalt von Aufgabentypen und ein Wechsel zwischen Repräsentationen wird als Qualitätsmerkmal bewertet. **Für** die Lernwirksamkeit verschiedener Darstellungs- und Interaktionsformen sprechen mehrere Studien, die andeuten, dass eine Instruktion im ITS dann effektiver ist, wenn sie die Schüler*innen anleitet, 1) **eigene Ideen zu finden**, 2) mit verschiedenen anderen Ideen zu **vergleichen** und 3) **nach Kriterien zu sortieren**, anstatt sie sofort zur richtigen Antwort zu führen (vgl. Gerard et al., 2015). Alternative Darstellungsformen können Kindern mit Dyskalkulie helfen, einfacher einen Zugang

zu mathematischen Aufgaben zu finden (vgl. De Smedt et al., 2013 sowie Rauscher et al., 2016 und Kohn et al., 2020 für die Wirksamkeit von *Calcularis*). Des Weiteren führt, laut Ma et al. (2014), ein **Mix aus prozeduralem und deklarativem Wissen** zu den höchsten Lerneffekten ($g = 0.65$), wobei sich aber eine einförmige Mischung aus Textaufgaben und Multiple Choice als eher ungünstig erweist ($g = 0.06$). Für variable Aufgabentypen spricht auch, dass sich Kompetenzen zwischen Zahlenstrahltrainings und Vergleichsaufgaben nicht vollständig übertragen (vgl. Maertens et al., 2016).

Gegen die (sehr) variable Verwendung von Interaktionswerkzeugen und Aufgabentypen sprechen zum einen die bereits oben angeführten Gründe der kognitiven Überforderung. Darüber hinaus zeigen Ran et al. (2021) eine leichte Gegenevidenz: Systeme, die mehrere Arten von Fertigkeiten adressieren, hatten schwächere Effekte ($d = 0.2$) im Vergleich zu gezielten Systemen (problem solving: $d = 0.95$; math reasoning: $d = 1.18$; arithmetic skill: $d = 0.58$).

***bettermarks* und Wettbewerber:** Die Gegenüberstellung der Eingabemöglichkeiten und Aufgabentypen vergleichbarer Mathe-Lernsysteme zeigt, dass nur *ASSISTMENTS* und *ALEKS* lediglich Multiple Choice sowie Textfelder als Eingabeform vorhalten, dabei aber Bilder in den Aufgabenstellungen verwenden. *IXL* bietet zwar mehrere Aufgabentypen mit jeweils eigener Darstellungsform, dies aber eher eingeschränkt. *Cognitive Tutor* verwendet mehrere Darstellungsformen und Eingabewerkzeuge für Geometrie und lässt Algebra-Aufgaben schrittweise (symbolisch) bearbeiten. *Calcularis* schließlich scheint am ehesten mit *bettermarks* vergleichbar zu sein, da das – nachgewiesenen lernwirksame – System verschiedene Darstellungsformen (Zahlenstrahl, symbolisch, ikonisch) und reichhaltige Interaktionsformen mit diesen Darstellungen nutzt und darüber hinaus eine spielerische Aufbereitung verfolgt. Letzterer Ansatz lässt sich auch bei *DreamBox* (ebenfalls verschiedene Darstellungs- und interaktionsformen nutzend) beobachten. Zusammenfassend spricht die Analyse der Wettbewerber leicht für die Wirksamkeit. Zwei von drei Systemen mit reichhaltigen Darstellungen und Interaktionswerkzeugen sind nachweislich wirksam, während zwei von drei Systemen mit eher einfachen Darstellungen und Interaktionswerkzeugen schwächere Wirkungsevidenz zeigen.

Fazit: Wenngleich nur wenige Forschungsergebnisse den bei *bettermarks* eingeschlagenen Weg des Einsatzes von intelligenten Interaktionswerkzeugen in Kombination mit formulierten Lernzielen direkt unterstützen: Es spricht viel dafür, dass verschiedene Darstellungs- und Interaktionsformen lernförderlich sind.

B3. Zusatzfrage: Wie lernwirksam sind Gamification-Elemente?

Anzahl Befunde: 4

Kernbefunde: Für die Wirksamkeit von Gamification-Elementen sprechen folgende Befunde: Bai et al. (2020) listen vier theoretische Vorteile von Gamification-Elementen auf: a) **Enthusiasmus** erzeugen, b) **Leistungsfeedback** geben, c) **Wertschätzung** zeigen und d) **Zielsetzungsverhalten fördern** (d.h.: Lernende werden motiviert, sich eigene Ziele zu setzen und zu erreichen). Empirisch stellen sie in einer Metastudie einen insgesamt positiven Effekt von Gamification fest ($g = 0.54$). Auch Faghihi et al. (2014) zeigen, dass eine spielerische/gamifizierte Mathe-Lernumgebung wirksamer ist als *ALEKS*.

Gegen die Wirksamkeit von Gamification-Elementen sprechen folgende Befunde: Theoretisch geben Bai et al. (2020) an, dass Gamification zu **Versagensangst und Eifersucht** führen kann, wenn Wettbewerbselemente eingebaut werden. Empirisch zeigen Long und Aleven (2014), dass eine

spielerische Lernumgebung, die Lernende für Wiederholungen belohnt, schlechter abschneidet als ein herkömmliches ITS. Benavides-Varela et al. (2020) finden in ihrer Metastudie keine Unterschiede in der Wirksamkeit von Mathematik-Lernspielen und einfachen digitalen Tutorials oder Drillansätzen.

***bettermarks* und Wettbewerber:** Die Analyse der Wettbewerber ergibt **keine eindeutige Aussage zur Wirksamkeit von Gamification-Elementen**. Von den analysierten Plattformen sind *DreamBox* und *Calcularis* am ehesten spielerisch aufbereitet, haben aber sehr unterschiedliche Wirkungsevidenz. ALEKS, IXL, *Cognitive Tutor* und *ASSISTMENTS* enthalten wenig bis keine Gamification-Elemente und sind ebenfalls unterschiedlich wirksam.

Fazit: Die Evidenz zu Gamification-Elementen ist gemischt. Eine mögliche Erklärung sind die unterschiedlichen Funktionsweisen von Gamification-Elementen: Elemente, die eher von den Lerninhalten ablenken und das Spiel in den Vordergrund treten lassen (z.B. über Wettbewerbsmechanismen) können der Lernwirksamkeit schaden, während Elemente, die eher der Transparenz dienen (z. B. Münzen und Sterne als Erfolgsindikatoren) der eigenen Selbstwirksamkeitswahrnehmung und Lernmotivation nutzen können.

2.3 Kerneigenschaft C: Mikro-Adaptivität

Mikro-Adaptivität bezeichnet die adaptive Hilfestellung während einzelner Lernaktivitäten (vgl. Speroni, 2021). Insofern entspricht Mikro-Adaptivität dem „inner loop“ eines ITS nach VanLehn (2006). Der einfachste Fall von Mikro-Adaptivität ist die Rückmeldung, ob eine Aufgabe richtig oder falsch bearbeitet wurde. Darüber hinaus zeigt *bettermarks* Feedback für typische Fehler an, gibt Lernenden die Möglichkeit, ihre Antwort zu korrigieren, zeigt bei wiederholten Falschantworten die richtige Lösung an und verweist auf hilfreiche, weitere Lernressourcen (wie vorgerechnete Beispiele oder Glossare). Im Folgenden führen wir die Evidenz für jede dieser Strategien auf.

C1. Fehlerbasiertes Feedback ist lernwirksamer als einfaches Richtig/Falsch-Feedback.

Anzahl Befunde: 10

Kernbefunde: Für die Wirksamkeit sprechen folgende Befunde: Kulik und Fletcher (2016) definieren Systeme, die nur einfaches richtig/falsch-Feedback geben als Computer Aided Instruction (CAI) im Gegensatz zu vollwertigen ITS. CAI zeigen generell kleinere Lerneffekte als ITS, was indirekt für ausführlicheres Feedback spricht. Akin (2022) sowie Hillmayr et al. (2020) machen in ihrer Metastudie die gleiche Feststellung. Belland et al. (2016) untersuchen die Wirksamkeit von unterschiedlichen Arten von Scaffolding und finden für alle Arten positive Effektstärken ($g = 0.41$ oder größer) im Vergleich zu einfachen Kontrollen. Ran et al. (2021) vergleichen diverse Mathematik-Lernsysteme und finden **größere Effektstärken für ITS mit adaptivem Feedback** ($d = 0.8$ und größer) **als für reine Übungssysteme** ($d = 0.22$). Prihar et al. (2022) finden in einer Metastudie größere Wirkung für elaboriertes Feedback im Vergleich zu richtig/falsch-Feedback, wobei die größten Effekte für Scaffolding in Videoform gemessen werden ($d = 0.267$). Van der Kleij (2015) finden in ihrer Metastudie einen Zusatzeffekt von elaboriertem Feedback über reines richtig/falsch-Feedback hinaus ($d = 0.54$) und sogar über richtig/falsch-Feedback mit richtiger Lösung hinaus ($d = 0.39$).

Gegen die Wirksamkeit sprechen folgende Befunde: Shute und Rahimi (2017) argumentieren aus theoretischer Sicht, dass einfache Wissensmessungs-Systeme, die nur richtig/falsch-Feedback geben, bereits lernwirksam sein können, während **elaboriertes Feedback Lernende eher überfordern könne**

(sie zitieren aber auch Studien, die eine höhere Wirksamkeit elaborierteren Feedbacks belegen). Empirisch finden Ma et al. (2014) für Systeme, die Feedback geben, keine größeren Effektstärken als für Systeme, die kein Feedback geben. Selent und Heffernan (2015) finden keinen zusätzlichen Lerneffekt von fehlerbasiertem „push“-Feedback im Vergleich zu „pull“-Hinweisen (also Hinweisen, die Lernende auf eigenes Klicken hin aufdecken können).

bettermarks und Wettbewerber: Aus der Analyse der Wettbewerber lässt sich schließen, dass **umfassenderes Feedback eher mit Wirksamkeit korreliert**. Alle besonders wirksam evaluierten Systeme – *Cognitive Tutor*, *ASSISTMENTS* und *Calcularis* – geben detaillierteres Feedback als nur richtig/falsch, während die weniger wirksam evaluierten Systeme – *ALEKS*, *DreamBox* und *IXL* – vor allem richtig/falsch-Feedback geben.

Fazit: Insgesamt ergibt sich ein Bild, das deutlich **für** die Wirksamkeit ausführlichen Feedbacks spricht. Trotzdem ist es aus theoretischer Sicht geboten, Hilfestellungen nur dann bereit zu stellen, wenn sie gebraucht werden. Zum einen, um die Lernenden nicht mit zu vielen Informationen zu überfordern (Cognitive load-Theorie), zum anderen um sie zur eigenständigen Bearbeitung der Aufgaben zu führen (Scaffolding-Theorie).

C2. Die Möglichkeit zur Korrektur und Reflexion eigener Antworten ist lernwirksam.

Anzahl Befunde: 8

Kernbefunde: Wir konnten nur eine Studie finden, die direkt die Möglichkeit zur Korrektur nach einer Falschantwort untersucht: In einer umfassenden ($n = 2769$) Wirkungsstudie zum *ASSISTMENTS*-System wurde *ASSISTMENTS* als Hausaufgabenersatz in 43 Schulen eingesetzt (Murphy et al., 2020). Das *ASSISTMENTS*-System erlaubt Lernenden, ihre Antworten zu korrigieren und war – als Gesamtsystem – nachweislich wirksam ($g = 0.22$), was **indirekt für die Wirksamkeit der Korrekturmöglichkeit** spricht.

Umfassendere Forschung gibt es zur Wirksamkeit der Reflexion eigener Antworten. Thurm und Graewert (2020) sowie VanLehn (2011) argumentieren aus theoretischer Sicht für Reflexionsmöglichkeiten in ITS. Belland et al. (2016) finden in ihrer Metastudie **Wirksamkeitsevidenz für alle Formen elaborierten Scaffoldings** im Vergleich zu einfachen Kontrollen ($g = 0.41$ oder größer). Das schließt insbesondere metakognitives Scaffolding ($g = 0.42$) ein und damit auch Aufforderungen, die eigenen Antworten zu reflektieren. Verschaffel et al. (2019) stellen in ihrer Metastudie ebenfalls die Lernwirksamkeit metakognitiver Hinweise fest. Aleven und Koedinger (2010) weisen in ihrer Metastudie nach, dass die Reflexion der eigenen Lösungsschritte lernwirksam sein kann. Van der Kleij et al. (2015) stellen in ihrer Metastudie fest, dass **aufgabenbezogenes, regulatives Feedback** (also z.B. die Aufforderung, die eigene Antwort zu reflektieren und zu korrigieren) **besonders wirksam** war ($d = 1.05$ oder größer). Gerard et al. (2015) weisen in einer Metastudie die Wirksamkeit von umfassendem Feedback nach ($g = 0.27$), das zur Reflexion anregen soll, verglichen mit richtig/falsch-Feedback.

bettermarks und Wettbewerber: Die Analyse der Wettbewerber deutet ebenfalls darauf hin, dass die Möglichkeit zur Korrektur und Reflexion eigener Aufgaben mit Wirksamkeit korreliert. Die besonders wirksamen Systeme *ASSISTMENTS*, *Cognitive Tutor* und *Calcularis* lassen nach unseren Informationen eine Korrektur und Reflexion zu, während weniger wirksame Systeme wie *IXL* nur einen Versuch zulassen.

Fazit: Insgesamt gibt es wenig Studien, die direkt die Möglichkeit zur Korrektur untersuchen. Viel positive Evidenz gibt es aber dafür, Lernende zur Reflexion ihrer Antwort aufzufordern. Wenn die

Korrekturmöglichkeit also in eine pädagogische Strategie eingebunden ist, ist sie wahrscheinlich lernwirksam.

C3. Das Anzeigen von Lösungswegen ist eher lernwirksam.

Anzahl Befunde: 5

Kernbefunde: Bei wiederholt fehlerhafter Bearbeitung einer Aufgabe lässt das *bettermarks*-System Lernende auch ohne korrekte Lösung fortfahren, zeigt aber ein vorgerechnetes Lösungsbeispiel mit ausführlicher Kommentierung („worked example“) an.

Verschiedene Evidenz spricht **für** diese Strategie. Zunächst behandeln wir das Voranschreiten bei mehrfach fehlerhafter Lösung. Aus theoretischer Sicht argumentieren Baker et al. (2008), dass ein „Hängenbleiben“ bei einer Aufgabe zu erhöhter Frustration führt und Lernende eher zu unproduktivem Lernverhalten („gaming the system“) motiviert. Um solch ein „Hängenbleiben“ zu vermeiden kann es also sinnvoll sein, Lernende fortfahren zu lassen. Sales und Pane (2019) schließen aus Nutzungsdaten von *Cognitive Tutor*, dass Lernende, die trotz wiederholt inkorrekt Lösung weiter im Curriculum voranschreiten dürfen, eher mehr lernen als Lernende, die gezwungen werden, bei einem Thema zu verbleiben, bis sie es korrekt bearbeiten.

Weitere Evidenz liegt zu „worked examples“ vor. Das Studium solcher „worked examples“ kann nach Reed et al. (2012) ähnlich wirksam sein wie der *Cognitive Tutor*. Salden et al. (2009) weisen in einer Studie nach, dass „worked examples“ **mit anderer Mikro-Adaptivität kombiniert** werden können (so wie in *bettermarks*) und dann die **Lernwirkung noch erhöhen**.

Gegen das Anzeigen von Lösungswegen spricht aus didaktischer Sicht, dass es Lernenden die Möglichkeit gibt, absichtlich falsch zu antworten, um an eine richtige Lösung zu gelangen und schneller voranzuschreiten („gaming the system“, vgl. Baker et al., 2008). Allerdings diskutieren Baker et al. (2008), dass der Treiber für solches Verhalten nicht die Hilfsfunktion an sich ist, sondern eher Frustration mit der usability der Plattform oder der Schwierigkeit der Aufgaben. Sie empfehlen daher, usability und Aufgaben-Schwierigkeit zu verbessern, statt die Hilfsfunktion zu entfernen. Aus empirischer Sicht stellt die Metastudie von Van der Kleij et al. (2015) fest, dass Systeme, die die richtige Lösung anzeigen, nicht effektiver sind als Systeme mit einfachem richtig/falsch-Feedback. Allerdings ist dies lediglich korrelative Evidenz und betrachtet nicht das Anzeigen kommentierter richtiger Lösungen („worked examples“).

***bettermarks* und Wettbewerber:** Aus unserer Untersuchung der Wettbewerber wird nicht klar, welche Systeme Lösungswege anzeigen. Daher liegt uns hierzu keine Evidenz vor.

Fazit: Obwohl die Evidenz gemischt ist, sprechen die Befunde eher **für** das Anzeigen von Lösungswegen – zumindest, wenn die Lösungswege mit ausführlichem Feedback versehen sind und Lernende die vorgerechnete Lösung auch studieren. Selbst wenn Lernende das nicht tun, ist die *bettermarks*-Vorgehensweise aber sinnvoll, um ein frustrierendes „Hängenbleiben“ bei einer Aufgabe zu vermeiden.

C4. Zusätzliche Ressourcen (wie vorgerechnete Beispiele, Anleitungen und Videos) sind lernwirksam.

Anzahl Befunde: 11

Kernbefunde: Bei zusätzlichen Ressourcen sprechen wir vor allem von „pull“-Feedback, also Hilfestellungen, die Lernende aktiv vom System erfragen, während C1-C3 vor allem „push“-Feedback behandelt haben, das Lernende vom System proaktiv angeboten bekommen.

Umfangreiche Evidenz spricht **für** die Wirksamkeit zusätzlicher Ressourcen. Die Arbeiten von Reed et al. (2012) und Salden et al. (2009) sprechen etwa für vorgerechnete Beispiele (siehe C3). Gong et al. (2012) fanden heraus, dass Lernende, die **zusätzliche Websites** zu einem Thema aufrufen, bei Folgeaufgaben zum gleichen Thema im Mittel besser abschnitten als Lernende, die das nicht taten. Allerdings handelt es sich um eine Observationsstudie, kein kontrolliertes Experiment. Schwonke et al. (2013) zeigen, dass die Lernwirkung des *Cognitive Tutor* für **schwächere Lernende** steigt, wenn Lernenden eine **zusätzliche Anleitung** für das System bereitgestellt wird und Long und Alevén (2013) zeigen die **Lernwirkung eines „Skill-Tagebuchs“**, in dem Lernende mit Hilfe der Plattform ihren Fortschritt nachhalten können. Leite et al. (2022) weisen nach, dass auch ein System, das ausschließlich **Videos** empfiehlt, lernwirksam sein kann. Schließlich gibt es auch qualitative Hinweise, dass die in *bettermarks* bereit gestellten Zusatzressourcen wirksam sind. Im Rahmen der Wirkungsstudie von Scharnagl et al. (2014) zu *bettermarks* wurden auch Freitextkommentare der Lernenden erhoben, in denen die zusätzlichen Erklärungen, die *bettermarks* bereitstellt, besonders gelobt wurden.

Gegen die Wirksamkeit weiterer Ressourcen spricht lediglich ein Befund von Hillmayr et al. (2020), nämlich dass Hypermedien allein wenig Lerneffekt zeigen. Dieser Befund studiert allerdings nicht die Kombination von Hypermedien mit anderen Inhalten und Mikro-Adaptivität, wie es in *bettermarks* der Fall ist.

Mehrere Befunde geben außerdem Hinweise für die optimale Ausgestaltung weiterer Ressourcen. Aus theoretischer Sicht ist zu beachten, dass es auch lernförderlich sein kann, Informationen *nicht* zu geben (vergleiche „**Assistance Dilemma**“ nach Koedinger und Alevén, 2007). Köller et al. (2022) warnen insbesondere davor, dass **zu viele Hilfestellungen Lernende kognitiv überfordern** können. Aus Scaffolding-Sicht ist es geboten, Hilfestellungen nur dann bereit zu stellen, wenn sie wirklich benötigt werden, um Lernende an die eigenständige Lösung heranzuführen (Hammond und Gibbons, 2019). Auch die Studie von Salden et al. (2009) legt nahe, dass es für die Wirksamkeit entscheidend ist, Ressourcen bedarfsgerecht und adaptiv bereit zu stellen.

Walkington et al. (2015) fanden in ihrer großangelegten Studie zu Zusammenhängen der Lesbarkeit und der Themen mathematischer Textaufgaben und den Lösungserfolgen mehrere Lesbarkeitsindikatoren und Themen der Aufgabenstellungen, die mit dem Lernerfolg zusammenhängen: So führen **schwierige, mehrdeutige und unkonkrete Wörter** zu mehr Hilfeanfragen insbesondere leistungsschwächerer Schüler*innen. Gleiches Verhalten hängt des Weiteren mit der **Textlänge** zusammen: eine Algebra-Textaufgabe mit **drei Sätzen** kann erfolgreicher gelöst werden als mit vier und mehr Sätzen. Sie sollten darüber hinaus eine gewisse **Satzähnlichkeit** (hinsichtlich Semantik, Länge, Wörter) aufweisen, denn mit mehreren sehr unterschiedlichen Sätzen gehen auch eine geringere Genauigkeit und mehr Hilfeanfragen einher. Die Verwendung der dritten Person Einzahl als Präposition war in der Studie hingegen mit höherer Genauigkeit und weniger Hilfeanfragen korreliert. Vor dem Hintergrund, dass *bettermarks* häufig Alltagsbeispiele verwendet, ist von

Bedeutung, dass sich beim Lösen von Algebra-Textaufgaben eine hohe **Vertrautheit mit dem Thema** positiv auswirkt: Eine geringere Genauigkeit der Lösungen zeigte sich bei Arbeits- bzw. Finanzthemen, eine bessere Genauigkeit bei Beispielen aus dem Privat- und Sozialleben. Gleichwohl ist bei leseschwächeren Kindern die Mehrdeutigkeit von Wörtern ausschlaggebender als die Themenvertrautheit.

***bettermarks* und Wettbewerber:** Die Analyse der Wettbewerber gibt keinen eindeutigen Hinweis für die Lernwirksamkeit zusätzlicher Ressourcen. Die meisten Plattformen – nachweislich wirksam oder nicht – stellen ein gewisses Maß zusätzlicher Ressourcen bereit, von einfachen Links auf Vorlesungsmaterial bis hin zu aufgabenspezifischen Hinweisen (im Falle von *Cognitive Tutor* und *ASSISTMENTS*, zum Beispiel). Systeme mit aufgabenspezifischen Zusatzmaterialien, die adaptiv bereitgestellt werden, scheinen allerdings tendenziell lernwirksamer zu sein.

Fazit: Insgesamt spricht die Evidenz **für** die Wirksamkeit zusätzlicher Ressourcen. Allerdings betonen mehrere Quellen, dass solche Ressourcen adaptiv bereitgestellt werden sollten, um Lernende nicht zu überfordern und dass die einfache Sprache und usability der Ressourcen entscheidende Bedeutung hat.

2.4 Kerneigenschaft D: Makro-Adaptivität

Wir definieren Makro-Adaptivität als die Auswahl von Lerninhalten und -aktivitäten gemäß des individuellen Lernfortschritts (siehe auch „outer loop“ von VanLehn, 2006, und die Erklärung von Speroni, 2021). In *bettermarks* wird die Makro-Adaptivität einerseits durch Lehrkräfte sichergestellt, die Inhalte angepasst für einzelne Lernende vergeben können. Das System kann aber auch automatisch Wissenslücken erkennen (wenn bestimmte Fehler wiederholt auftreten) und Übungen vorschlagen, um diese Wissenslücken zu schließen. Diese automatische Strategie ist Gegenstand unserer Untersuchung.

D1. Das Identifizieren von Wissenslücken sowie die Vergabe gezielter Übungen, um diese Wissenslücken zu schließen, hat keinen eindeutigen Effekt.

Anzahl Befunde: 9

Kernbefunde: Zunächst stellen wir fest, dass kein Befund direkt die Wirksamkeit der Strategie von *bettermarks* beforcht. Am nächsten kommt die Metastudie von Ma et al. (2014). Dort werden Systeme, die „Misconceptions“ (also Fehlvorstellungen, in etwa vergleichbar mit Wissenslücken) modellieren, verglichen mit Systemen, die das nicht tun. Es wird kein Lernwirksamkeitsunterschied festgestellt. Allerdings ist diese Evidenz rein korrelativ und behandelt vor allem Systeme, die mit *bettermarks* kaum vergleichbar sind.

Allerdings gibt es zahlreiche Befunde, die Makro-Adaptivität in weiterem Sinne behandeln – also das Vergeben von Aufgaben gemäß dem Wissens- und Fertigungsstand der Lernenden. Aus theoretischer Sicht argumentieren Koedinger und Alevan (2007) sowie Ritter et al. (2007) für Makro-Adaptivität, die sie als Umsetzung der Theorie des „**Mastery Learning**“ verstehen (vgl. Guskey, 2010). Außerdem könne Makro-Adaptivität die Zeit der Lernenden optimal nutzen, weil **nur Lernaktivitäten unternommen würden, die noch zur Erweiterung von Wissen und Fertigkeiten beitragen**. Köller et al. (2022) argumentieren aus gesellschaftlicher Perspektive für Makro-Adaptivität: So könnten besondere **Förderbedarfe** erkannt und damit Bilder inklusiver gestaltet werden. Thurm und Graewert (2022)

fordern Diagnostik von Kompetenzen und Wissenslücken aus mathematik-didaktischer Sicht. VanLehn (2011) beschreibt diese **Diagnostik als eine der großen Stärken von ITS**, die menschliche Lehrkräfte aus Mangel an Zeit an Daten nicht gleichermaßen leisten können. Zusammenfassend ist Makro-Adaptivität aus theoretischer Perspektive eindeutig zu empfehlen.

Empirisch ist die Wirksamkeit von Makro-Adaptivität allerdings bisher kaum untersucht – vermutlich, weil es schwierig ist, zwei Systeme zu bauen, die in allen Belangen vergleichbar sind außer im Hinblick auf Makro-Adaptivität. Am nächsten kommen die folgenden drei Befunde. Murphy et al. (2020) finden, dass innerhalb von ASSISTMENTS die Nutzung stärker makro-adaptive Lerninhalte („SkillBuilders“) eher mit Lernerfolg korreliert als die Nutzung anderer Inhalte. Allerdings zeigen Maghfur et al. (2009) in einem kontrollierten Experiment (n=87) **keinen zusätzlichen Lerneffekt von Mastery Learning** in ASSISTMENTS. Schließlich stellen Sales und Pane (2019) fest, dass Lernende in *Cognitive Tutor*, die trotz mangelnder „mastery“ zum nächsten Thema übergehen, schneller lernen als Lernende, die Themen erst komplett meistern. Dieser Befund spricht eher gegen Makro-Adaptivität, ist aber keine kausale Evidenz.

bettermarks und Wettbewerber: Viele Wettbewerber definieren die **Makro-Adaptivität** – im Sinne einer Wissensdiagnostik und darauf basierender Auswahl von Aufgaben – **als Kern ihrer Systeme**. Allerdings sind die Ansätze zur Makro-Adaptivität in allen Plattformen unterschiedlich – und keiner der Ansätze ist direkt mit *bettermarks* vergleichbar. *Cognitive Tutor* versucht, über Bayesian Knowledge Tracing die „Mastery“ der Lernenden feingranular zu messen und ist insgesamt lernwirksam (Ritter et al., 2007). ASSISTMENTS modelliert „Mastery“ über ein sehr simples Schema: Drei richtige Antworten zu einem Thema werden als indikativ für „Mastery“ bewertet. ASSISTMENTS ist insgesamt nachweislich lernwirksam, allerdings spricht die Studie von Maghfur et al. (2009) dafür, dass die Makro-Adaptivität daran vielleicht keinen großen Anteil hat. *Calcularis* und *ALEKS* setzen unterschiedliche Varianten von Wissensgraphen ein, um Rückschlüsse über den Wissens- und Fertigungsstand der Lernenden zu führen und damit die nächsten Aufgaben auszuwählen. *Calcularis* ist damit wirksam, während die Evidenz für *ALEKS* gemischt ist. *Area9* wirbt damit, das Tempo durch die Lerninhalte zu beschleunigen, wenn Lernende richtige Antworten geben, aber wir konnten keinerlei Lernwirkungsevidenz zu *Area9* finden. *DreamBox* wirbt damit, „true adaptive learning“ anzubieten, bei dem nicht nur das Tempo durch das Curriculum angepasst wird, sondern gänzlich unterschiedliche Wege eröffnet würden. Das System hat allerdings keine eindeutige Wirkungsevidenz.

Insgesamt spricht die Analyse der Wettbewerber dafür, dass nur bestimmte Varianten von Makro-Adaptivität mit Lernwirkung korrelieren. Insbesondere sind *Cognitive Tutor* und *Calcularis* hervorzuheben, die beide feingranular den Lernfortschritt modellieren. ASSISTMENTS zeigt allerdings auch, dass ein System ohne ausgefeilte Makro-Adaptivität durchaus lernwirksam sein kann.

Fazit: Die pädagogische Theorie spricht eindeutig für Makro-Adaptivität: Idealerweise sollten Aufgaben gerade so ausgewählt werden, dass sie dem Fertigkeiten- und Kenntnisstand der Lernenden entsprechen (siehe Mastery Learning und Zone of proximal development). Allerdings scheinen nur wenige System diese Makro-Adaptivität erfolgreich automatisieren zu können (insbesondere *Cognitive Tutor* und *Calcularis*). Das spricht dafür, dass nur bestimmte Spielarten der (automatisierten) Makro-Adaptivität lernwirksam sind. Der Ansatz von *bettermarks* – Wissenslücken zu erkennen und dann gezielt zu schließen – ist noch nicht durch Wirkungsstudien untersucht worden. Hier besteht also eine **Forschungslücke**.

Trotz der gemischten Evidenzlage **könnte eine automatisierte Modellierung des Lernfortschritts den *bettermarks*-Ansatz unterstützen** – in dem der Lernfortschritt an Lernende und Lehrende kommuniziert wird und damit Lehrende makro-adaptive Entscheidungen treffen können. Die Umfragen mit Lehrkräften in *ASSISTMENTS* durch Gillespie et al. (2022) und die Befunde von Spitzer und Musslick (2021) bei der Auswertung von *bettermarks*-Logs sprechen für einen solchen Ansatz.

2.5 Rahmenbedingungen E: Einsatz im Klassenverbund und Abdeckung des Curriculums

Eine zentrale Frage im Zusammenhang mit dem Einsatz digitaler Lerntools ist die nach dem generellen Mehrwert gegenüber traditionellem, analogem Unterricht. *bettermarks* ist auf die überwiegende Nutzung im Nachmittags-, d. h. Hausaufgabenbereich ausgelegt. In welchem Ausmaß und in welchen Einsatzszenarien – durch Lehrkraft begleiteter Einsatz, Einsatz im Klassenverbund, Nutzung in Einzel- oder Gruppenarbeit – ist Gegenstand der Literaturanalyse, deren Kernergebnisse im Folgenden vorgestellt werden. Ergänzt werden die Ergebnisse durch einen erneuten Blick auf Befunde zur Lernwirksamkeit bei vollständiger Abdeckung des Curriculums (wie bei *bettermarks*, siehe auch Kapitel 2.1, A2.).

E1. Der Einsatz von ITS mit Beteiligung einer Lehrkraft ist im Vergleich zu traditionellem, nicht-digitalen Unterricht eher lernwirksam, wenn die Lehrkraft ausreichend geschult ist.

Anzahl Befunde: 17

Kernbefunde: Zahlreiche Befunde belegen eine **Lernwirksamkeit von Intelligenten Tutoring-Systemen**, welche die des traditionellen, nicht digitalen Unterrichts übertrifft. Akin (2022) findet in ihrer Metastudie zur Wirksamkeit von web-basierten Mathe-Lernplattformen **positive Effekte für alle Klassenstufen**, schließt aber vor dem Hintergrund der Reihenfolge der Effektgrößen – absteigend: 1. Studierende, 2. Mittelschule, 3. Gymnasien, 4. Grundschule –, dass Erwachsene bzw. **ältere Lernende noch mehr** profitieren von web-basierten Mathe-Instruktionen als Kinder, da sie über bessere Selbstkontrollfähigkeiten, digitale Fähigkeiten, mathematische Vorkenntnisse und motivierende Überzeugungen verfügen.

Allgemein sprechen die Forschungsergebnisse **für einen durch die Lehrkraft begleiteten Einsatz des ITS** (z. B. Sahni et al., 2021; Ran et al., 2021). In ihrer Wirkungsstudie zu *bettermarks* fanden Scharnagl et al. (2014) einen dosisabhängigen Effekt: Schüler*innen, deren Lehrer*innen **mehr Übungsreihen** einsetzten, profitierten stärker als die Schüler*innen, deren Lehrer*innen nur einige wenige einsetzten. Zur Frage nach der lernwirksamsten Nutzungsintensität fassen Ran et al. (2021) mehrere Meta-Studien zusammen, die zeigen, dass kürzere Interventionen erfolgreicher verlaufen als längere, konnten dieses Ergebnis in der eigenen Studie allerdings nicht bestätigen.

Wenn ITS in den regulären Unterricht integriert sind oder als Aktivität nach der Schule bzw. für Hausaufgaben genutzt werden, hat das nach Ma et al. (2014) einen positiven Effekt ($g = 0.37$). Dabei sei der Einsatz von **ITS in Einzelarbeit wirksamer** als die herkömmliche Instruktion **in großen Gruppen** ($g = 0.44$). Allerdings ist die Lernwirksamkeit von ITS in Einzelarbeit ununterscheidbar von Kleingruppenlernen und **leicht schlechter als menschliches 1zu1-Tutoring**. Auch Kulik und Fletcher (2016) zeigen in ihrer Meta-Analyse zur Wirksamkeit von ITS, dass diese effektiver sind als 'conventional instruction' in Gruppen ($d = 0.37$) und weniger effektiv als 'human tutoring' ($d = -0.25$).

Generell würde der Effekt menschlichen Tutorings aber überschätzt und Effektstärken hängen entscheidend mit der Art der Evaluierung zusammen (siehe dazu die Erläuterungen zur Wirkungsforschung zu digitalen Lernumgebungen und Herausforderungen in Metastudien im Abschnitt 1.2 des Berichts).

Besonders ausschlaggebend – so zeigen neun Befunde – sind die **Kompetenzen und Einbindung der Lehrkräfte bei der Einführung eines ITS**. Eine Rolle spielt zum einen der **Erfahrungseffekt**: Lehrkräfte mit ITS-Erfahrung erzeugen bei Lernenden einen deutlich höheren Lerneffekt ($d = 0.96$) als ITS-unerfahrene Lehrkräfte ($d = -0.23$), zeigten bereits Koedinger und Anderson 1993 (zitiert nach Kulik und Fletcher, 2016). Dieser Effekt wird durch eine der größten Studien zur Effektivität von *Cognitive Tutor* mit mehr als 13.000 Lernenden (Pane et al., 2013) bestätigt: Es zeigte sich keine Wirkung im ersten, dafür aber im zweiten Jahr der Benutzung des Systems im tatsächlichen Schulunterricht, da die Lehrer*innen ihr **Verhalten angepasst** hatten und **geübter im Umgang mit dem System** waren.

Darüber hinaus wirken sich Erfahrungen mit dem ITS-Einsatz nicht nur auf den Umgang mit dem System, sondern **generell auf die Unterrichtsgestaltung** aus: Eine Längsschnittstudie mit mehreren Fallbeispielen zum Einsatz von *ASSISTMENTS* durch Mathematiklehrer der 7. Klasse und zu den Auswirkungen auf Unterricht vor und während der COVID-19-Pandemie von Gillespie und Kollegen (2022) zeigte, dass Lehrkräfte ihre Unterrichtspraktiken zur Überprüfung der Hausaufgaben und zur Feststellung, ob ihre Schüler*innen den Unterricht verstanden haben, in allen Unterrichtskontexten (persönlich, Fernunterricht und beides kombiniert) geändert haben. Sie nutzten zum einen automatisch erstellte Berichte, um die Hausaufgabenkontrolle auf der Grundlage der Leistungen der Schüler*innen zu fokussieren und pädagogische Interventionen und/oder Nachhilfeunterricht anzubieten. Zum anderen wurden Unterrichtsvorschläge der *ASSISTMENTS*-Plattform für die Unterrichtsstundenplanung bei Stoffwiederholungen gut angenommen.

Die Einbindung und Schulung der Lehrkräfte bereits bei der Einführung von ITS wird explizit angeraten, insbesondere in Deutschland (vgl. Thurm und Graewert, 2020). Wenn die Lehrer*innen über geringe Kompetenzen für personalisiertes Lernen verfügen und nicht ausreichend über die Systeme und ihre Möglichkeiten informiert und im Einsatz trainiert werden, kommt es zu schlechteren Bewertungen und Ergebnissen, zeigte eine große Studie zur Anwendung von *ALEKS* als komplementäres Werkzeug zur Klassenraum-Lehre: Nur sechs Prozent der Klassen schafften es, *ALEKS* richtig anzuwenden und mit dem System personalisierte Lehre durchzuführen (vgl. Phillips u. a., 2020).

***bettermarks* und Wettbewerber:** Der – lediglich korrelative – Vergleich mit den Wettbewerbern spricht ebenfalls für die Wirksamkeit von ITS, die Lehrkräfte stark einbinden. Die beiden am besten validierten Systeme – *ASSISTMENTS* und *Cognitive Tutor* – setzen beide auf besonders deutliche Einbindung der Lehrkräfte, während Systeme mit weniger ausgeprägter Wirkungsevidenz eher als unabhängiges Selbstlernwerkzeug oder Hausaufgabenwerkzeug genutzt werden (*ALEKS*, *DreamBox*, *IXL*). Die einzige Ausnahme ist *Calcularis*: Dieses System wird als Förderinstrument zusätzlich zum regulären Schulunterricht empfohlen und bedarf daher keiner direkten Einbindung der Lehrkraft.

Fazit: Der Einsatz von ITS in Begleitung oder zumindest mit Anleitung durch eine Lehrkraft im Unterricht und für das Erledigen von Hausaufgaben ist lernwirksamer als traditioneller Mathe-Unterricht. Eine Voraussetzung dafür scheinen die ausreichende Schulung und vorhandene Erfahrungen und Kompetenzen der Lehrenden mit digitalen Lerntools allgemein und den konkreten ITS zu sein. Durch die Verwendung von ITS verändert sich nachweislich die Unterrichtsgestaltung,

unabhängig davon, ob der Unterricht in Präsenz (in Gruppenarbeit), zu Hause (in Einzelarbeit) oder im Wechsel hybrid stattfindet.

E2. = A3. Eine höhere Abdeckung eines Curriculums korreliert nicht eindeutig mit Lernwirksamkeit, ist aber als Rahmenbedingung, um im deutschen Markt bestehen zu können, unabdingbar.

Eine weitere auf empirische Befunde hin untersuchte Rahmenbedingung ist gleichzeitig eine zentrale Systemeigenschaft von *bettermarks*: die vollständige Abdeckung des Curriculums. Diese Eigenschaft wurde bereits im Rahmen der Forschungsfragen zur Kerneigenschaft A „Vollständigkeit der Aufgaben und Inhalte“ auf Befunde zur Lernwirksamkeit betrachtet. Daher wird an dieser Stelle noch einmal das Fazit der diesbezüglichen Literaturanalyse wiedergegeben.

Fazit: Insgesamt spricht viel dafür, dass ein System mit einer breiteren Abdeckung des Curriculums grundsätzlich zu bevorzugen ist. Insbesondere erfordert der Schulkontext eine breite Abdeckung von Inhalten, damit das System auch über weite Teile eines Schuljahres einsetzbar bleibt (statt nur für einzelne Stunden und Themen). Es scheint aber **ein Spannungsverhältnis zwischen breiter Abdeckung von Inhalten und gleichzeitig hoher Aufgabenqualität** zu bestehen. Systeme, die sich eher für die Breite entscheiden (wie *ALEKS*, *Dreambox*, *IXL* und *ASSISTMENTS*) haben oft eine geringe Wirkung, während eher eng gefasste Systeme (wie *Cognitive Tutor* und *Calcularis*) in ihren jeweiligen Bereichen wirksam sind. Es gilt also, bei einer Ausdehnung der Inhalte eine weiterhin hohe Qualität der Aufgaben zu gewährleisten.

3 Empfehlungen für die weitere Ausgestaltung des *bettermarks*-Systems

Aus den empirischen Befunden lassen sich Empfehlungen für die zukünftige Weiterentwicklung von *bettermarks* schließen, insbesondere für den deutschen Markt. Dabei unterscheiden wir Empfehlungen, die das System betreffen und Empfehlungen, die die Anwendung des Systems betreffen.

3.1 Empfehlungen zur Systemgestaltung

1. Den Fokus auf Mikro-Adaptivität beibehalten und ausbauen (Eigenschaft C)

Ein überraschendes Ergebnis unserer Untersuchung ist, dass wir insgesamt stärkere Evidenz für Mikro-Adaptivität im Vergleich zu Makro-Adaptivität finden konnten, das heißt: Die Hilfestellung während der Bearbeitung einzelner Aufgaben ist entscheidender als die (automatische) Auswahl der Aufgaben. Dies deckt sich mit der derzeitigen Strategie von *bettermarks*: **Im Vergleich zu Wettbewerber-Systemen ist bei *bettermarks* die Mikro-Adaptivität stark, die Makro-Adaptivität aber weniger stark ausgeprägt. Dieser Fokus sollte beibehalten werden.**

Trotzdem lassen sich für die Mikro-Adaptivität noch Empfehlungen ableiten, die über das derzeitige System hinaus gehen. Studien legen insbesondere nahe, dass Lernende von **metakognitiven und regulativen Hilfestellungen** profitieren, das heißt: Lernende sollten durch geeignete Hilfestellungen befähigt werden, den eigenen Lernstand einzuschätzen, sich selbst Ziele zu setzen und zu erreichen und strategisches Wissen über das eigene Lernen aneignen. Derlei metakognitive Hilfestellungen und Erklärungen könnten noch verstärkt eingesetzt werden.

Ferner deutet die Metastudie von Prihar et al. (2022) darauf hin, dass Hilfestellungen in **Videoformat** lernwirksamer sind als Texte. *bettermarks* könnte also erwägen, besonders umfangreiche Hilfestellungen auf ein Videoformat umzustellen

Schließlich ist bei der Mikro-Adaptivität aber stets die pädagogische Theorie zu beachten: Es gilt, Hilfestellungen nur dann zu geben, wenn Lernende nicht eigenständig weiterkommen (Scaffolding; Hammond und Gibbons, 2019) und die Lernenden durch Anzahl und Umfang der Hilfestellungen nicht zu überfordern (cognitive load). In anderen Worten: **Es kann lernwirksamer sein, Hilfestellungen nicht zu geben, als sie zu geben** (Assistance Dilemma, vgl. Koedinger und Aleven, 2007). Daher ist die Adaptivität der Hilfestellungen entscheidend.

2. Makro-Adaptivität in den Händen der Lernenden und Lehrenden belassen aber durch automatisierte Werkzeuge unterstützen (Eigenschaft D)

Obwohl viele Wettbewerber gerade mit der Makro-Adaptivität ihrer Systeme werben, fällt die Wirkungsevidenz zu Makro-Adaptivität erstaunlich uneindeutig aus. Insgesamt deutet die Evidenz darauf hin, dass (gut informierte) Lehrkräfte mindestens ebenso gute makro-adaptive Entscheidungen treffen können wie automatisierte Systeme. **Dies stützt die bisherige Strategie von *bettermarks*, die Auswahl der Übungsinhalte vor allem den Lehrkräften zu überlassen.** Diese Strategie sollte – auch angesichts der besonders hohen Ansprüche an menschliche Autonomie in der EU und in Deutschland – beibehalten werden.

Trotzdem kann es sich für *bettermarks* lohnen, **eine automatisierte Diagnose des Lernfortschritts** („knowledge tracing“) zu ermöglichen, nämlich um **Lernenden und Lehrenden Auskunft über den**

Lernfortschritt zu geben und so menschliche Adaptivitäts-Entscheidungen zu unterstützen. Die Arbeit von Gillespie et al. (2022) deutet daraufhin, dass Lehrkräfte, die durch ein ITS über den Lernstand informiert sind, anders unterrichten können als Lehrkräfte, die diese Information nicht haben.

3. Hohe Qualität und Granularität der Inhalte bewahren (Eigenschaft A)

Die Wirkungsevidenz spricht für eine feine Auflösung von Aufgabenschritten und Lernzielen. Daher spricht vieles dafür, **auch bei der Gestaltung der zukünftigen Lerninhalte in *bettermarks* auf solch feine Granularität und insgesamt hohe Aufgabenqualität Wert zu legen.** Bei begrenzten Ressourcen steht dieses Ziel allerdings im Spannungsverhältnis mit dem Ziel, viele Inhalte bereit stellen zu können. Daher ist anzumerken, dass **die Breite der Inhalte und ihre Anzahl allein nicht eindeutig mit höherer Lernwirksamkeit assoziiert** ist. Gleichwohl kann es durch andere Rahmenbedingungen (siehe unten) geboten sein, neue Inhalte zu erstellen. Wir empfehlen lediglich, die Breite der Inhalte nicht *für sich selbst genommen* als lernwirksam zu betrachten.

4. Interaktionswerkzeuge beibehalten (Eigenschaft B)

Unsere Untersuchung zeigt, dass reichhaltige und vielfältige Interaktionswerkzeuge generell lernwirksam sind. **Der Fokus von *bettermarks* auf geeignete Werkzeuge sollte daher beibehalten werden.** Die Literatur gibt keine eindeutigen Hinweise darauf, dass *bettermarks* in diesem Bereich Werkzeuge fehlen, daher besteht hier kein Handlungsbedarf.

3.2 Empfehlungen zu Rahmenbedingungen

1. Die Einbindung der Lehrkräfte verstärken (Rahmenbedingung E)

Wirkungsstudien geben starke Hinweise darauf, dass die Einbindung der Lehrkräfte ein Schlüssel ist, um Mathematik-Lernumgebungen im Schulalltag wirksam zu machen. Insbesondere müssen Lehrkräfte die pädagogische Strategie der Systeme verstehen und in ihre eigene Pädagogik einbetten. Dazu braucht es Fortbildung und Erfahrung. Pane et al. (2016) zeigen in einer großangelegten Wirkungsstudie, dass das *Cognitive Tutor*-System erst im zweiten Jahr Wirkung zeigte, als sich der Betrieb des Systems eingespielt hatte. Murphy et al. (2020) berichten, dass die großflächige Wirkungsstudie des *ASSISTMENTS*-Systems mit Lehrkräfte-Fortbildungen vorbereitet wurde – dann ließ sich ein Effekt zeigen ($g = 0.22$). Kulik und Fletcher (2016) zitieren Vorarbeiten, die darauf hindeuten, dass Klassen mit system-erfahrenen Lehrkräften einen hohen Lernerfolg haben, während Klassen mit system-unerfahrenen Lehrkräften überhaupt keinen Lerneffekt zeigen. Auch in *bettermarks*-bezogenen Studien gibt es Evidenz, dass das System besonders dann wirksam ist, wenn Lehrkräfte es strategisch einzusetzen wissen (Spitzer et al., 2021).

Insgesamt ergibt sich aus der Evidenz die eindringliche Empfehlung, einen Schwerpunkt der *bettermarks*-Strategie (weiterhin) auf die Einbindung der Lehrkräfte zu legen. Lehrkräfte sollten in **Fortbildungen** nicht nur den Umgang mit der Plattform als Anwender*innen lernen, sondern vor allem, das System in die jeweils eigene **pädagogische Strategie** einzubetten. Dafür kann es auch hilfreich sein, *bettermarks* mit weiteren Unterrichtsinhalten auszugestalten (über Übungsaufgaben hinaus), sodass es Lehrkräften leichter fällt, *bettermarks*-Übungsaufgaben und den Unterricht im Klassenraum in Einklang zu bringen. Eine weitere Faustregel geben Kulik und Fletcher (2016): **Bei erfahrenen Lehrkräften steht das ITS und dessen Benutzeroberfläche im Hintergrund und die Pädagogik im**

Vordergrund. Lehrkräfte sollten so vertraut mit *bettermarks* werden, dass es ein natürliches Werkzeug der Unterrichtsgestaltung wird – wie ein Schulbuch jetzt. Solch ein Lehrkräfte-orientiertes Paradigma dient auch der Akzeptanz durch Lehrkräfte und der Transparenz des Systems.

2. Inhalte für den curricularen Bedarf ausbauen (Eigenschaft A)

Um Lehrkräften zu ermöglichen, *bettermarks* während des gesamten Schuljahres unterrichtsunterstützend einzusetzen, müssen zu allen wesentlichen Teilen des jeweiligen Curriculums Inhalte in *bettermarks* zur Verfügung stehen. Obwohl also die Breite der Inhalte nicht für sich genommen lernwirksam ist, **erfordert der schulische Einsatz eine breite Abdeckung des Curriculums**. Trotzdem ist zu empfehlen, bestehende Inhalte möglichst effizient zu nutzen. Daher ist es erforderlich, **Inhalte für Lehrkräfte einfach durchsuchbar** zu machen, etwa indem Lehrenden ermöglicht wird, die *bettermarks*-Inhalte gemäß ihres jeweiligen Landes- oder Schulcurriculums zu sortieren.

4 Forschungsdesiderata und -potenziale für gemeinsame Follow-up-Projekte

Die Analyse der empirischen Befunde hat einige Forschungsdesiderata und – zumindest für den deutschsprachigen Raum – offene Forschungsfragen aufgezeigt, welche Potenzial für gemeinsame Studien oder Follow-up-Projekte bieten. Wir unterscheiden Wirkungsstudien für die Weiterentwicklung von *bettermarks*, Wirkungsstudien zum Einsatz im landesweiten Vergleich sowie eine gemeinsame Forschungsstrategie:

Wirkungsstudien und Forschungsfragen für die Weiterentwicklung von *bettermarks*

- Während aus pädagogischer Perspektive alles für **Makro-Adaptivität**, also die automatische Zuweisung von zu Kenntnisstand und Lernerkompetenzen passenden Aufgaben spricht, liegen bislang nur uneindeutige empirische Ergebnisse zu ihrer Lernwirksamkeit vor. Vor diesem Hintergrund wäre eine **Wirkungsstudie zum Schließen von Wissenslücken** sinnvoll.
- Während die Wirksamkeit reichhaltiger und vielfältiger **Interaktionswerkzeuge** insgesamt belegt ist, ist die Evidenz für die spezifischen Werkzeuge, die bei *bettermarks* eingesetzt werden, noch dürftig. Da die Qualität und Quantität der Werkzeuge ein Alleinstellungsmerkmal von *bettermarks* ist, sollte deren Wirkung systematisch(er) untersucht und belegt werden. Insbesondere vor dem Hintergrund einer möglichen kognitiven Überlastung durch eine zu große Darstellungs- und Interaktionsvielfalt.

Wirkungsstudien zum *bettermarks*-Einsatz im landesweiten Vergleich

- In Anknüpfung an frühere Forschungsergebnisse zur Wirkung von *bettermarks* schlagen wir eine Wirkungsstudie zum Vergleich des Lernerfolgs bei Hausaufgaben mit *bettermarks* vs. ohne *bettermarks* unter Kontrolle des Zeitaufwands für die Hausaufgaben, für mehr Themen als nur Brüche und mit Unterscheidung nach Leistungsniveaus und Mathe-Angst vor (Scharnagl et al., 2014, hatten gezeigt, dass schwächere Lernende weniger profitieren). Solch eine Studie ließe sich sowohl für Berlin als auch im landesweiten Vergleich durchführen.

- Ebenso ließe sich über die Resultate in landesweiten Vergleichsarbeiten die Wirkung unterschiedlicher Nutzungsintensitäten und Einsatzszenarien von *bettermarks* systematisch untersuchen. Mögliche Forschungsfragen wären hier:
 - Wie intensiv und systematisch muss die Nutzung/der Einsatz von *bettermarks* für eine größtmögliche Lernwirksamkeit gestaltet sein?
 - Welche Auswirkungen hat der Einsatz verstehensorientierten Lernmaterials von *bettermarks* im Unterricht auf den außerunterrichtlichen Übungsbedarf?

Gemeinsame Forschungsstrategie *bettermarks* – EdTec

- Inspiriert von *ASSISTMENTS*, das A/B-Tests auf der eigenen Plattform durchführt (vgl. Prihar et al., 2022), regen wir an, auch in der *bettermarks*-Infrastruktur die Durchführung kontrollierter Studien zu ermöglichen, um die **kontinuierliche, evidenzbasierte Weiterentwicklung von *bettermarks* und den wirksamsten Einsatzszenarien** zu unterstützen. Hier könnten in Zusammenarbeit mit dem EdTec Lab in standardisierten Testdesigns z. B. einzelne Plattform-Funktionen, Hilfestellungen, Interaktionswerkzeuge u.v.m. systematisch getestet werden.
- In einer solchen gemeinsamen Forschungsstrategie könnten auch weitere von Scharnagl et al. (2014) formulierte Forschungsfragen verfolgt werden: Warum führt *bettermarks* zu höherem Lernerfolg? Welche Rolle spielen die Kerneigenschaften Mikro- und Makro-Adaptivität? Haben reichhaltige Eingabewerkzeuge und vielfältige Darstellungs- und Interaktionsformen vor allem motivationale/affektive Effekte oder inwiefern tragen sie zur Lernwirksamkeit bei?
- Schließlich ließe sich – mit Blick auf noch ausstehende empirische Evidenz zum Einfluss von Moderatorvariablen wie das Geschlecht, die Art der Veröffentlichung, die Lernansätze, die Lernmaterialien, die Dauer der Durchführung und das Jahr, oder auch sozioökonomische Merkmale (Akin et al. 2022) – die zukünftige Forschung gewinnbringend gestalten.

Abbildungsverzeichnis

Abbildung 1: Übersicht der Fundstellen.....	9
---	---

Tabellenverzeichnis

Tabelle 1: Suchbegriffe.....	8
Tabelle 2: In- und Exklusionskriterien	8
Tabelle 3: Übersicht der Wirkungsstudien zu Wettbewerbern.....	10

Literaturverzeichnis

- Akin, A. (2022): The effectiveness of web-based Mathematics instruction (WBMI) on K-16 students' mathematics learning: a meta-analytic research. *Education and Information Technologies*, 27, S. 8015-8040. <https://doi.org/10.1007/s10639-022-10931-x>
- Aleven, V.; Koedinger, K. (2010): An effective metacognitive strategy: learning by doing and explaining with a computer-based *Cognitive Tutor*. *Cognitive Science*, 26(2), S. 147-179. https://doi.org/10.1207/s15516709cog2602_1
- Arms, D. (2019): Middle School Mathematics, Student Growth, and the Role of Technology-assisted, Independent Practice. Dissertation. Ball State University, Muncie, IN, USA. https://cardinalscholar.bsu.edu/bitstream/handle/123456789/201653/ArmsD_2019-2_BODY.pdf
- Bai, S.; Hew, K.; Huang, B. (2020): Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, 30, Artikel 100322. <https://doi.org/10.1016/j.edurev.2020.100322>
- Baker, R.; Walonoski, J.; Heffernan, N.; Roll, I.; Corbett, A.; Koedinger, K. (2008): Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research*, 19(2), S. 185-224. <https://www.learntechlib.org/primary/p/24328/>
- Barton, J. (2016): Mathematics Intervention Utilizing Carnegie Learning's *Cognitive Tutor*® and Compass Learning's Odyssey Math®. Dissertation, Wilmington University, DE, USA.
- Basadr, T. (2020): Effectiveness of the *Cognitive Tutor* on Mathematics Achievement of High School Students. Dissertation. Baker University. Baldwin City, KS, USA. https://www.bakeru.edu/images/pdf/SOE/EdD_Theses/Basadre_Theresa.pdf
- Belland, B.; Walker, A.; Kim, N.; Lefler, M. (2016). Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis. *Review of Educational Research*, 87(2), S. 309-344. <https://doi.org/10.3102/0034654316670999>

- Benavides-Varela, S.; Callegher, C.; Fagiolini, B.; Leo, I.; Altoè, G.; Lucangeli, D. (2020): Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers & Education*, 157, Artikel 103953. <https://doi.org/10.1016/j.compedu.2020.103953>
- BIXLer, S. (2019): One-to-one iPad Technology in the Middle School Mathematics and Science Classrooms. *International Journal of Technology in Education and Science*, 3(1), S. 1-18. <https://www.learntechlib.org/p/207269/>
- Brown, A. (2018): How Does Online Practice in a Public, Middle School Mathematics Classroom Affect Mathematical Understanding? Dissertation. Trevecca Nazarene University. Nashville, TN, USA. <https://www.proquest.com/openview/c2ddd5b8502115f831bae887e8615137/1>
- Danso, R. (2017): The Impact of Increased Technology Integration on the Achievement of Students. Master's thesis. Goucher College. Towson, MD, USA. <http://hdl.handle.net/11603/4370>
- De Smedth, B.; Noël, M.; Gilmore, C.; Ansari, D. (2013): How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, 2(2), S. 48-55. <https://doi.org/10.1016/j.tine.2013.06.001>
- Donnelly, S.; Parmar, R. (2022): A Comparison of *IXL* Math Online Software and Textbook Problems for Homework In An Urban Middle School. *Journal of Computers in Mathematics and Science Teaching*, 41(1), S. 45-67. <https://www.learntechlib.org/primary/p/219062/>
- Faghihi, U.; Brautigam, A.; Jorgenson, K.; Martin, D.; Brown, A.; Measures, E.; Maldonado-Bouchard, S. (2014): How Gamification Applies for Educational Purpose Specially with College Algebra. *Procedia Computer Science*, 41, S. 182-187. <https://doi.org/10.1016/j.procs.2014.11.102>
- Fang, Y.; Ren, Z.; Hu, X.; Graesser, A. (2018): A meta-analysis of the effectiveness of *ALEKS* on learning, *Educational Psychology*, 39(10), S. 1278-1292, <https://doi.org/10.1080/01443410.2018.1495829>
- Fanusi, A. (2015): The effect of *ALEKS* math support on standardized math test scores in middle school. Dissertation. Liberty University. Lynchburg, VA, USA.
- Gardes, M.; Hugli, C.; Dewi, J.; Hanssen, L.; Deruaz, M. (2022): Evaluation of a Computer-based Learning Program for Students with Mathematical Learning Difficulties. In: Bolondi, G.; Ferretti, F. (Hg.) Tagungsband des 12th congress of the European Society for Research in Mathematics Education (CERME). <https://hal.science/hal-03745424/>
- Gerard, L.; Matuk, C.; McElhaney, K.; Linn, M. (2015): Automated, adaptive guidance for K-12 education. *Educational Research Review*, 15, S. 41-58. <https://doi.org/10.1016/j.edurev.2015.04.001>
- Gillespie, J.; Winn, K.; Faber, M.; Hunt, J. (2022). Implementation of a Mathematics Formative Assessment Online Tool Before and During Remote Learning. In: Rodrigo, M.; Matsuda, N.; Cristea, A.; Dimitrova, V. (Hg.). Tagungsband der 23rd International Conference on Artificial Intelligence in Education (AIED). Teil II: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium, S. 168-173. https://doi.org/10.1007/978-3-031-11647-6_29

- Gong, Y.; Beck, J.; Heffernan, N. (2012): WEBSistments: Enabling an Intelligent Tutoring System to Excel at Explaining Rather Than Coaching. In: Cerri, S.; Clancey, W.; Papadourakis, G.; Panourgia, K. (Hg.). Tagungsband der 11th International Conference on Intelligent Tutoring Systems (ITS), S. 268-273. https://doi.org/10.1007/978-3-642-30950-2_34
- Grams, D. (2018): A Quantitative Study of the Use of *DreamBox* Learning and Its Effectiveness in Improving Math Achievement of Elementary Students with Math Difficulties. Dissertation. Northcentral University. San Diego, CA, USA.
- Guskey, T. (2010): Lessons of mastery learning. *Educational Technology*, 68(2), S. 52-57. https://uknowledge.uky.edu/edp_facpub/14
- Hammond, J.; Gibbons, P. (2019): What is scaffolding? In: Burns, A.; de Silva Joyce, H. (Hg.): Teachers' voices 8: Explicitly supporting reading and writing in the classroom. Macquarie University. Sydney, NSW, Australia. <https://www.dr-hatfield.com/educ341/readassign/1%20explicitly%20supporting%20reading%20&%20writing.pdf#page=15>, S. 8-16.
- Hillmayr, D.; Ziernwald, L.; Reinhold, F.; Hofer, S.; Reiss, K. M. (2020): The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153, Artikel 103897, <https://doi.org/10.1016/j.compedu.2020.103897>
- Hu, X.; Craig, S.; Bargagliotti, A.; Graesser, A.; Okwumabua, T.; Anderson, C.; Cheney, K.; Sterbinsky, A. (2012): The Effects of a Traditional and Technology-based After-school Setting on 6th Grade Student's Mathematics Skills. *Journal of Computers in Mathematics and Science Teaching*, 31(1), S. 17-38. <https://www.learntechlib.org/primary/p/38628/>
- Kerner, M. (2016): An Analysis of the Effectiveness of *ALEKS* within a High School Mathematics Setting. Dissertation. Aurora University. Aurora, IL, USA.
- Koedinger, K.; Aleven, V. (2007): Exploring the Assistance Dilemma in Experiments with *Cognitive Tutors*. *Educational Psychology Review* 19, S. 239-264. <https://doi.org/10.1007/s10648-007-9049-0>
- Kohn, J.; Rauscher, L.; Kucian, K.; Käser, T.; Wyszkon, A.; Esser, G.; von Aster, M. (2020): Efficacy of a Computer-Based Learning Program in Children With Developmental Dyscalculia. What Influences Individual Responsiveness? *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.01115>
- Köller, O.; Thiel, F.; Akeren, I.; Anders, Y.; Becker-Mrotzek, M.; Cress, U.; Diehl, C.; Kleickmann, T.; Lütje-Klose, B.; Prediger, S.; Seeber, S.; Ziegler, B.; Kuper, H.; Stanat, P.; Maaz, K.; Lewalter, D. (2022): Digitalisierung im Bildungssystem: Handlungsempfehlungen von der Kita bis zur Hochschule. Gutachten der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz (SWK). peDOCS. <https://doi.org/10.25656/01:25273>
- Kulik, J.; Fletcher, J. (2016): Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1), S. 42-78. <https://doi.org/10.3102/0034654315581420>

- Leite, W.; Roy, S.; Chakraborty, N.; Michailidis, G.; Huggins-Manley, A. C.; D’Mello, S.; Faradonbeh, M. K. S.; Jensen, E.; Kuang, H.; Jing, Z. (2022): A novel video recommendation system for algebra: An effectiveness evaluation study. In: Wise, A.; Martinez-Maldonado, R.; Hilliger, I. (Hg.). Tagungsband der 12th International Learning Analytics and Knowledge Conference (LAK), S. 294-303. <https://doi.org/10.1145/3506860.3506906>
- Long, Y.; Aleven, V. (2013): Skill Diaries: Improve Student Learning in an Intelligent Tutoring System with Periodic Self-Assessment. In: Lane, H. C.; Yacef, K.; Mostow, J.; Pavlik, P. (Hg.). Tagungsband der 16th International Conference on Artificial Intelligence in Education (AIED), S. 249-258. https://doi.org/10.1007/978-3-642-39112-5_26
- Long, Y.; Aleven, V. (2014): Gamification of Joint Student/System Control over Problem Selection in a Linear Equation Tutor. In: Trausan-Matu, S.; Boyer, K.E.; Crosby, M.; Panourgia, K. (Hg.). Tagungsband der 12th International Conference on Intelligent Tutoring Systems (ITS), S. 378-397. https://doi.org/10.1007/978-3-319-07221-0_47
- Ma, W.; Adesope, O.; Nesbit, J.; Liu, Q. (2014): Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), S. 901–918. <https://doi.org/10.1037/a0037123>
- Maertens, B.; De Smedt, B.; Sasanguie, D.; Elen, J.; Reynvoet, B. (2016): Enhancing arithmetic in pre-schoolers with comparison or number line estimation training: Does it matter? *Learning and Instruction*, 46, S. 1-11. <https://doi.org/10.1016/j.learninstruc.2016.08.004>
- Maghfur, A.; Zhong, B.; Gamache, C. (2009): Mastery Learning in Cognitive Computer Tutors. Forschungsbericht. Worcester Polytechnic Institute, MA, USA. <https://core.ac.uk/download/pdf/212976335.pdf><https://digitalcommons.wpi.edu/iqp-all/1197>
- Murphy, R.; Roschelle, J.; Feng, M.; Mason, C. (2020): Investigating Efficacy, Moderators and Mediators for an Online Mathematics Homework Intervention, *Journal of Research on Educational Effectiveness*, 13(2), S. 235-270. <https://doi.org/10.1080/19345747.2019.1710885>
- Palculict, R. (2022): The Effects of Computer-Assisted Instruction on Student Learning of Fractions in Middle School Mathematics. Dissertation. Liberty University. Lynchburg, VA, USA. <https://digitalcommons.liberty.edu/doctoral/3823/>
- Pane, J.; Griffin, B.; McCaffrey, D.; Karam, R. (2016). Effectiveness of *Cognitive Tutor Algebra I* at Scale. *Educational Evaluation and Policy Analysis*, 36(2), S. 127-144. <https://doi.org/10.3102/0162373713507480>
- Pane, J., McCaffrey, D.; Slaughter, M.; Steele, J.; Ikemoto, G. (2010): An Experiment to Evaluate the Efficacy of *Cognitive Tutor Geometry*. *Journal of Research on Educational Effectiveness*, 3(3), S. 254-281. <https://doi.org/10.1080/19345741003681189>
- Pellegrini, M.; Lake, C.; Neitzel, A.; Slavin, R. (2021): Effective Programs in Elementary Mathematics: A Meta-Analysis. *AERA Open*, 7. <https://doi.org/10.1177/2332858420986211>

- Phillips, A.; Pane, J.; Reumann-Moore, R.; Shenbanjo, O. (2020): Implementing an adaptive intelligent tutoring system as an instructional supplement. *Educational Technology Research and Development*, 68, S. 1409–1437. <https://doi.org/10.1007/s11423-020-09745-w>
- Prihar, E.; Syed, M.; Ostrow, K.; Shaw, S.; Sales, A.; Heffernan, N. (2022): Exploring Common Trends in Online Educational Experiments. In: Cristea, A.; Brown, C.; Mitrovic, A.; Bosch, N. (Hg.): Tagungsband der International Conference on Educational Data Mining (EDM). <https://doi.org/10.5281/zenodo.6853041>
- Rafi, A.; Anuar, K.; Samad, A.; Hayati, M., & Mahadzir, M. (2005): Improving spatial ability using a web-based virtual environment (WbVE). *Automation in Construction*, 14(6), S. 707–715. <https://doi.org/10.1016/j.autcon.2004.12.003>
- Ran, H.; Kim, N. J.; Secada, W. G. (2021): A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in K-12 classrooms. *Journal of Computer Assisted Learning*, 38(1), S. 258–284. <https://doi.org/10.1111/jcal.12611>
- Rauscher, L.; Kohn, J.; Käser, T.; Mayer, V.; Kucian, K.; McCaskey, U.; Esser, G.; von Aster, M. (2016): Evaluation of a Computer-Based Training Program for Enhancing Arithmetic Skills and Spatial Number Representation in Primary School Children. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00913>
- Reed, S.; Corbett, A.; Hoffman, B.; Wagner, A.; MacLaren, B. (2012): Effect of worked examples and *Cognitive Tutor* training on constructing equations. *Instructional Science*, 41, S. 1-24. <https://doi.org/10.1007/s11251-012-9205-x>
- Rholetter, K. (2020): A Causal-Comparative Study on the Efficacy of Intelligent Tutoring Systems on Middle-Grade Math Achievement. Dissertation. Liberty University. Lynchburg, VA, USA. <https://digitalcommons.liberty.edu/doctoral/2427/>
- Ritter, S.; Anderson, J.; Koedinger, K.; Corbett, A. (2007): *Cognitive Tutor*: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14, S. 249-255. <https://doi.org/10.3758/BF03194060>
- Rochelle, J.; Murphy, R.; Feng, M.; Bakia, M. (2017): How big is that? Reporting the Effect Size and Cost of ASSISTments in the Maine Homework Efficacy Study. *ASSISTments in the Maine Homework Efficacy Study. Forschungsbericht. SRI International. Menlo Park, CA, USA.* <https://digitalcommons.library.umaine.edu/mepri/139>
- Sabo, K.; Atkinson, R.; Barrus, A.; Joseph, S.; Perez, R. (2013): Searching for the two sigma advantage: Evaluating algebra intelligent tutors. *Computers in Human Behavior*, 29(4), S. 1833-1840. <https://doi.org/10.1016/j.chb.2013.03.001>
- Sahni, D.; Polanin, J.; Zhang, Q.; Michaelson, L.; Caverly, S.; Polese, M.; Yang, J. (2021): A What Works Clearinghouse Rapid Evidence Review of Distance Learning Programs. *Forschungsbericht. Insittute of Education Sciences, U.S. Department of Education, Washington, DC, USA.* <https://files.eric.ed.gov/fulltext/ED610886.pdf>

- Salden, R.; Aleven, V.; Renkl, A.; Schwonke, R. (2009): Worked Examples and Tutored Problem Solving: Redundant or Synergistic Forms of Support? *Topics in Cognitive Science*, 1(1), S. 203-213. <https://doi.org/10.1111/j.1756-8765.2008.01011.x>
- Sales, A.; Pane, J. (2019): The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, 13(1), S. 420-443. <https://doi.org/10.1214/18-AOAS1196>
- Sawilowsky, S. (2009): New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2), Artikel 26. <https://doi.org/10.22237/jmasm/1257035100>
- Scharnagl, S.; Evanschitzky, P.; Streb, J.; Spitzer, M.; Hille, K. (2014): Sixth Graders Benefit from Educational Software when Learning about Fractions: A Controlled Classroom study. *Numeracy*, 7(1), Artikel 4. <http://doi.org/10.5038/1936-4660.7.1.4>
- Schuetz, R.; Biancarosa, G.; Goode, J. (2018): Is Technology the Answer? Investigating Students' Engagement in Math. *Journal of Research on Technology in Education*, 50(4), S. 318-332. <https://doi.org/10.1080/15391523.2018.1490937>
- Schwonke, R.; Ertelt, A.; Otieno, C.; Renkl, A.; Aleven, V.; Salen, R. (2013): Metacognitive support promotes an effective use of instructional resources in intelligent tutoring. *Learning and Instruction*, 23, S. 136-150. <https://doi.org/10.1016/j.learninstruc.2012.08.003>
- Selent, D.; Heffernan, N. (2015): When More Intelligent Tutoring in the Form of Buggy Messages Does not Help. In: Conati, C.; Heffernan, N.; Mitrovic, A.; Verdejo, M. (Hg.) *Tagungsband der International Conference on Artificial Intelligence in Education (AIED)*, S. 768-771. https://doi.org/10.1007/978-3-319-19773-9_112
- Shute, V.; Rahimi, S. (2017): Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1), S. 1-19. <https://doi.org/10.1111/jcal.12172>
- Sottolare, R.; Graesser, A.; Hu, X.; Holden, H. (2013): Design recommendations for intelligent tutoring systems: Volume 1 – learner modelling. US Army Research Laboratory, Orlando, FL, USA. <https://play.google.com/store/books/details?id=2leEBAAQBAJ>
- Speroni, C. (2021). *Aus Fehlern lernen – Das Konzept hinter bettermarks*. Webseite. <https://de.bettermarks.com/konzept-bettermarks/>
- Spitzer, M.; Musslick, S. (2021): Academic performance of K-12 students in an online-learning environment for mathematics increased during the shutdown of schools in wake of the COVID-19 pandemic. *PLOS ONE*. <https://doi.org/10.1371/journal.pone.0255629>
- Steenbergen-Hu, S.; Cooper, H. (2013): A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), S. 970–987. <https://doi.org/10.1037/a0032447>
- Stein, M. K.; Grover, B. W. & Henningsen, M. (1996): Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms.

- American educational research journal, 33(2), S. 455–488. <https://doi.org/10.2307/1163292>
- Sun, S.; Else-Quest, N.; Hodges, L.; French, A.; Dowling, R. (2021): The Effects of *ALEKS* on Mathematics Learning in K-12 and Higher Education: A Meta-Analysis. *Investigations in Mathematics Learning*, 13(3), S. 182-196. <https://doi.org/10.1080/19477503.2021.1926194>
- Tempelaar, D.; Rienties, B.; Kaper, W.; Giesbers, B.; Schim van der Loeff, S.; van Gastel, L.; van de Vrie, E.; van der Kooij, H.; Cuypers, H. (2012): Mathematics Bridging Education Using an Online, Adaptive E-Tutorial: Preparing International Students for Higher Education. In: Juan, A.; Huertas, M.; Trenholm, S.; Steegmann, C. (Hg.). *Teaching Mathematics Online: Emergent Technologies and Methodologies*. IGI Global. Hershey, PA, USA. S. 167-186. <https://doi.org/10.4018/978-1-60960-875-0.ch008>
- Thurm, D.; Graewert, L. (2022): *Digitale Mathematik-Lernplattformen in Deutschland*. Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Deutschland. <https://doi.org/10.1007/978-3-658-37520-1>
- Van der Kleij, F.; Feskens, R.; Eggen, T. (2015): Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research*, 85(4), S. 475-511. <https://doi.org/10.3102/0034654314564881>
- VanLehn, K. (2006): The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 16(3), S. 227-265. <https://www.public.asu.edu/~kvanlehn/Stringent/PDF/06IJAIED.pdf>
- VanLehn, K. (2011): The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), S. 197-221. <https://doi.org/10.1080/00461520.2011.611369>
- Van Ruler, D. (2017): *Blended Learning and Math Growth: Investigating the Role of IXL Math in the Growth of 5th Grade Students' Math Fluency Scores*. Master's thesis. Dordt University. Sioux Center, IA, USA. https://digitalcollections.dordt.edu/med_theses/117/
- Verschaffel, L.; Depaepe, F.; Mevarech, Z. (2019): Learning Mathematics in Metacognitively Oriented ICT-Based Learning Environments: A Systematic Review of the Literature. *Educational Research International*, 2019, Artikel 3402035. <https://doi.org/10.1155/2019/3402035>
- Walkington, C.; Clinton, V.; Ritter, S. N.; Nathan, M. J. (2015): How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, 107(4), S. 1051–1074. <https://doi.org/10.1037/edu0000036>
- Wang, H.; Woodworth, K. (2011): *A Randomized Controlled Trial of Two Online Mathematics Curricula*. Forschungsbericht. Society for Research on Educational Effectiveness. Evanston, IL, USA. <https://eric.ed.gov/?id=ED528686>
- Yilmaz, B. (2017): *Effects of Adaptive Learning Technologies on Math Achievement: A Quantitative Study of ALEKS Math Software*. Dissertation. University of Missouri-Kansas City. Kansas City, MO, USA.

https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/60519/Dissertation_2017_Yilmaz.pdf?sequence=1