Rezeptions- und Interpretationsprozesse von Lehrpersonen bei datengestützten Entscheidungen.

Exploration und Förderung

Dissertation

zur Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Sarah Damaris Bez

Tübingen

2024

1. Betreuer:	Prof. Dr. Thorsten Bohl
2. Betreuer:	Prof. Dr. Marcus Syring
Tag der mündlichen Prüfung:	22.04.2024
Dekan:	Prof. Dr. Ansgar Thiel
1. Gutachter:	Prof. Dr. Thorsten Bohl
2. Gutachter:	Prof. Dr. Marcus Syring

Danksagung

Mein erster herzlicher Dank geht an Prof. Dr. Thorsten Bohl und Prof. Dr. Marcus Syring für die Betreuung und Begutachtung meiner Promotion und auch ausdrücklich für die Möglichkeit, auf der Basis einer Lehrstuhlstelle als wissenschaftliche Mitarbeiterin promovieren zu können. Weiterhin danke ich Prof. Dr. Taiga Brahm herzlich für die Übernahme des Prüfungsvorsitzes.

Der nächste aufrichtige Dank geht an Prof. Dr. Samuel Merk, der auch nach seinem Weggang aus der Abteilung Schulpädagogik diese Arbeit weiter intensiv begleitet und unterstützt hat und von dem ich die letzten Jahre sehr viel gelernt habe.

Zudem danke ich allen Lehrkräften und Lehramtsstudierenden sehr, die an den empirischen Studien, die dieser Arbeit zugrunde liegen, teilgenommen haben und sich auf das laute Denken eingelassen haben. Ohne sie wäre diese Arbeit nicht möglich gewesen.

Ein weiterer herzlicher Dank gilt allen Kolleg*innen der Abteilung Schulpädagogik, des Instituts für Erziehungswissenschaft, der Tübingen School of Education, der Pädagogischen Hochschule Karlsruhe sowie aus anderen Hochschulen für die inhaltlichen Anregungen und die kollegiale Unterstützung beim Promovieren an sich. Ein besonderes Dankeschön geht an Dr. Malte Ring für den spannenden Austausch und das intensive Korrekturlesen des Mantelteils.

Weiterhin danke ich allen Koautor*innen der zu dieser Arbeit gehörenden Publikationen für die gute Zusammenarbeit sowie allen Hilfskräften und Masterstudierenden, die an unterschiedlichen Stellen die Datenerhebung und -aufbereitung unterstützt haben.

Der letzte Dank geht von Herzen an alle Menschen aus meinem persönlichen Umfeld, die mich auf diesem Weg unterstützt, ermutigt und begleitet haben.

Zusammenfassung

Im Rahmen der sogenannten Neuen Steuerung im Bildungswesen wird von Lehrkräften erwartet, Schule und Unterricht auf der Basis von Daten zu gestalten und weiterzuentwickeln. Dies wird häufig mit dem Ziel (verbesserter) fachlicher Leistungen von Schüler*innen verknüpft. Im Zuge der Digitalisierung im Bildungswesen liegen vermehrt leicht zugängliche Daten vor, die Lehrkräfte für datengestützte Entscheidungen nutzen können. Allerdings ist die Verfügbarkeit von Daten nur eine notwendige aber nicht hinreichende Voraussetzung für gelingende datengestützte Entscheidungen: Rahmenmodelle konzeptualisieren datenbasierte Entscheidungen von Lehrpersonen als einen zyklisch-sequenziellen Prozess, in dem Daten zuerst rezipiert und interpretiert werden müssen, um daraus Schlussfolgerungen für Anschlusshandlungen ziehen zu können, die dann umgesetzt und in ihren intendierten und nicht intendierten Wirkungen überprüft werden können. Studien, die den Umgang von Lehrpersonen mit Daten untersuchen, deuten insgesamt darauf hin, dass die *data literacy* von Lehrpersonen, also die entsprechenden Kompetenzen, eher niedrig bis moderat ausgeprägt sind und es Lehrpersonen schwer zu fallen scheint, Daten für die Gestaltung und Entwicklung von Schule und Unterricht zu nutzen. Dabei gelten Studien, die die Datenrezeption und -interpretation von Lehrpersonen fokussieren, als Desiderat. Dies gilt insbesondere für Studien auf der Mikroprozessebene mit einem Schwerpunkt auf ökologischer Validität, also für Studien, die Aussagen über alltägliche Prozesse erlauben. Diese Arbeit untersucht vor diesem Hintergrund die Datenrezeption und -interpretation von Lehrpersonen erstens explorativ fokussiert auf alltägliche kognitive Prozesse, zweitens konfirmatorisch mit Blick auf die Förderung entsprechender Kompetenzen bereits bei Lehramtsstudierenden und drittens konzeptuell hinsichtlich der entsprechenden Kompetenzen von Lehrpersonen im Kontext der zunehmenden Digitalisierung.

Zur Adressierung von Forschungsfrage 1 wurden zwei explorative Studien durchgeführt (Artikel 1 und 2), in denen Lehrpersonen gebeten wurden, laut zu denken, d.h. ihre Gedanken und Überlegungen laut zu äußern, während sie aktuelle Leistungsdaten ihrer Klassen rezipierten und interpretierten. Dabei zeigte sich, dass die Lehrpersonen in der Tendenz die Daten mit niedriger bis mittlerer Komplexität rezipierten, d.h. direkt gegebene Werte ablasen oder miteinander in Beziehung setzten, wobei sich aber in beiden Studien deutliche Heterogenität zwischen den Lehrpersonen zeigte. Zudem lässt sich auf Basis von Studie 2 (Artikel 2) die Hypothese generieren, dass der Abgleich der Ergebnisse mit der eigenen Perspektive und die Analyse von Fehlern wichtig für die Formulierung von

Handlungsmaßnahmen sind. Forschungsfrage 2 wurde untersucht, indem eine Interventionsstudie mit Lehramtsstudierenden durchgeführt wurde (Artikel 3), in der die Datenrezeption und -interpretation in einem Onlinesetting gefördert werden sollte. Dabei zeigte sich ein großer positiver Effekt auf die Kompetenzen von Lehramtsstudierenden. Zur Adressierung von Forschungsfrage 3 wurde auf der Basis von konzeptuellen Modellen und bisherigen Forschungsergebnissen für die These argumentiert, dass die *data literacy* von Lehrpersonen eine notwendige Voraussetzung dafür ist, dass sich die Potenziale, die mit einer verstärkten Digitalisierung für datengestützte Entscheidungen einhergehen, entfalten und mögliche dysfunktionale Wirkungen minimiert werden können.

Auch wenn diese Arbeit nicht die gesamte (angenommene) Wirkungskette von datengestützten Entscheidungen untersucht, sondern lediglich die Rezeption und Interpretation von Daten fokussiert, trägt sie zu Erkenntnissen bei, wie Lehrkräfte in ihrem Alltag mit Daten umgehen, wie die entsprechenden Kompetenzen bereits bei angehenden Lehrpersonen gefördert werden und welche zentrale Rolle *data literacy* im Kontext der Digitalisierung einnimmt.

Summary

Teachers are supposed to use data as a basis for decisions concerning learning, instruction and developing schools, mainly to foster student learning. With the rise of digitization and technology, more data is more easily accessible. However, as conceptual frameworks indicate, data itself is not sufficient for improvement: As a first and crucial step, teachers have to make sense of data, i.e. notice and interpret them, to derive actions and finally evaluate these actions. Generally speaking, research reports that teachers show low to middle levels of data literacy and have difficulties concerning data-based decision making. However, there is more research needed on how teachers make sense of data. In particular, little is known about how teachers notice and interpret data in their daily practice from a process perspective. Against this background, this dissertation focuses on sensemaking and investigates, first, how teachers make sense of data in their daily practice, second, how data literacy of pre-service teachers can be fostered, and third, which role data literacy of teachers plays in the context of digitization.

Concerning research question one, two exploratory think-aloud studies were conducted (paper 1 and 2). Teachers were asked to verbalize their thoughts while they made sense of the latest assessment results of their students. As a general result of both studies, most teachers showed low to middle levels of complexity, although there was substantial variance among teachers. In addition, one can derive the hypothesis from study two that analyzing errors and comparing results with the teacher's personal perspective play an important role for constructing instructional implications. Concerning research question two, it was investigated how data literacy, especially sensemaking, of pre-service teachers can be fostered in an online setting (paper 3). This study shows a large positive effect on the data literacy of the students. To address research question three (paper 4), based on conceptual frameworks and previous research, it was argued that data literacy of teachers is a necessary prerequisite to unlock potentials and to minimize possible side effects of data-based decision making in schools in the context of digitization.

Although this dissertation does not focus on the whole process of data-based decision making but only on sensemaking, it contributes to research in the field of data-based decision making and provides insights how teachers make sense of data in their daily practice, how data literacy of pre-service teachers can be fostered effectively and which central role data literacy of teachers plays in the context of digitization.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Schule und Unterricht beständig weiterzuentwickeln und dabei auch Daten zu nutzen ist im Kontext der sogenannten Neuen Steuerung im Bildungswesen von zentraler Bedeutung (Altrichter et al., 2016): Daten, verstanden als eine Form von Evidenz, sollen von unterschiedlichen Akteursgruppen im Mehrebenensystem genutzt werden, um schulische und unterrichtliche Prozesse zu gestalten und weiterzuentwickeln (Dedering & Kallenbach, 2023). Dies wiederum soll zu verbesserten Outcomes führen, worunter häufig (höhere) fachliche Leistungen von Schüler*innen verstanden werden (Altrichter et al., 2016). In diesem Prozess sollen Lehrpersonen eine Schlüsselrolle übernehmen, da es zu ihren genuinen Aufgaben gehört, Schule und Unterricht weiterzuentwickeln (KMK, 2000). Zwar liegen durch die zunehmende Datafizierung und Digitalisierung vereinfacht und vermehrt Daten vor, aber das Vorhandensein von Daten ist nur eine notwendige und keinesfalls hinreichende Bedingung für die angenommene Wirkungskette (Schildkamp, 2019): Denn ein zentraler erster Schritt zu Beginn des skizzierten zyklisch-sequenziellen Prozesses besteht darin, dass Lehrpersonen die Daten (adäquat) rezipieren und interpretieren, bevor Schlüsse für mögliche Anschlusshandlungen gezogen werden, diese umgesetzt und in ihren Wirkungen überprüft werden können.

Studienergebnisse, die den Umgang von Lehrpersonen mit Daten fokussieren, deuten insgesamt darauf hin, dass deren entsprechende Kompetenzen eher niedrig ausgeprägt sind und Lehrpersonen die Nutzung von Daten in der Praxis als herausfordernd wahrnehmen (Mandinach & Schildkamp, 2021a). Hierbei liegt der methodische Schwerpunkt jedoch hauptsächlich auf quantitativen Fragebogenstudien in Form retrospektiver Selbstauskünfte (Altrichter et al., 2016), Interventionsstudien zur Förderung datengestützter Entscheidungen mit hoher interner Validität (Filderman et al., 2022) sowie (weniger) Interviewstudien (z.B. Goffin et al., 2023; Schliesing, 2017). Auf dieser Basis sind Aussagen mit Blick auf die alltäglichen Prozesse bezüglich der Nutzung von Daten, d.h. hinsichtlich der ökologischen Validität, nicht ohne weiteres möglich. Zudem gelten Studien mit einem Fokus auf die Datenrezeption und -interpretation, gerade auch auf einer Mikroprozessebene, als Desiderat (Goffin et al., 2022; Hebbecker et al., 2022; Schildkamp, 2019). Daher fokussiert die vorliegende Arbeit die Datenrezeption und -interpretation von Lehrpersonen als zentralen Schritt zu Beginn des Gesamtprozesses datengestützter Entscheidungen im schulischen Kontext. Die Arbeit untersucht die Datenrezeption und -interpretation von Lehrpersonen erstens explorativ fokussiert auf alltägliche kognitive Prozesse, zweitens konfirmatorisch mit Blick auf die Förderung entsprechender Kompetenzen bereits bei Lehramtsstudierenden und

drittens konzeptuell hinsichtlich der entsprechenden Kompetenzen von Lehrpersonen im Kontext der zunehmenden Digitalisierung.

Einführend werden grundlegende Begriffe, Modelle und Forschungsbefunde zu datenbasierten Entscheidungen im schulischen Kontext skizziert (Kapitel 2), bevor ein Überblick über das Konstrukt *data literacy* von Lehrpersonen gegeben wird und der Forschungsstand hierzu dargestellt wird (Kapitel 3). Daraufhin werden die schon in den vorherigen Kapiteln angebahnten Forschungsdesiderate gebündelt formuliert, um die Forschungsfragen der vorliegenden Arbeit abzuleiten und deren Adressierung in den durchgeführten Forschungsarbeiten zu skizzieren (Kapitel 4). Nach der Darstellung der durchgeführten Studien anhand der entsprechenden Artikel bzw. Manuskripte (Kapitel 5-8) erfolgen abschließend die Zusammenfassung und Diskussion der Ergebnisse, die Darstellung der Limitationen der Arbeit sowie ein Ausblick auf sich anschließende Forschungsfragen sowie die Formulierung möglicher Implikationen für die Praxis sowie die Aus- und Weiterbildung von Lehrpersonen.

2. Datenbasierte Gestaltung und Entwicklung von Schule und Unterricht

2.1. Konzeptuelle und begriffliche Klärungen

Nähert man sich der datengestützten Gestaltung und Entwicklung von Schule und Unterricht zunächst über bildungspolitische Rahmenbedingungen und Setzungen, lässt sich konstatieren, dass datenbasierte Schul- und Unterrichtsentwicklung in Deutschland seit den 2000er Jahren im Rahmen der sogenannten Neuen Steuerung zunehmend in den Fokus gerückt ist (Thiel et al., 2019; Wurster, 2019). Neben der Teilnahme an internationalen Schulleistungsstudien und der Einführung und Überprüfung von Bildungsstandards wurden verschiedene Verfahren zur datenbasierten Qualitätssicherung und -entwicklung in der Gesamtstrategie zum Bildungsmonitoring von der Kultusministerkonferenz (KMK) festgeschrieben (z.B. KMK, 2016b) und in den Ländern (teilweise unterschiedlich) implementiert (Thiel et al., 2019): Schulen insgesamt und insbesondere Lehrpersonen sollen Daten aus Vergleichsarbeiten, zentralen Abschlussprüfungen sowie internen und externen Evaluationen nutzen, um die Qualität von Schule und Unterricht gezielt weiterzuentwickeln. Dies gilt als Teil der sogenannten Evidenzbasierung im Bildungswesen (Altrichter et al., 2016). Auch wenn hierbei immer wieder auf begriffliche Unschärfen hingewiesen und unterschiedliche Verständnisse, was unter Evidenz subsumiert wird, thematisiert werden, werden im Allgemeinen Daten unter Evidenz subsumiert (z.B. Dedering & Kallenbach, 2023;

Groß Ophoff & Cramer, 2022; Zlatkin-Troitschanskaia, 2016). Auch international ist die Erwartung an Lehrkräfte und andere Akteursgruppen, als Teil evidenzinformierten Handelns (auch) Daten zu nutzen, weit verbreitet, auch wenn die dafür entwickelten Rahmenbedingungen unterschiedlich ausgestaltet sind (Brown & Malin, 2022). Dies betrifft etwa das Verhältnis von *accountability* (Rechenschaftslegung/Kontrolle) und *improvement* (Weiterentwicklung von Schule und Unterricht), weshalb *high stakes* von *low stakes* Kontexten unterschieden werden (Altrichter et al., 2016).

Im englischsprachigen Diskurs wird, wenn es um die datenbasierte Gestaltung und Entwicklung von Schule und Unterricht geht, von *data-based decision making* gesprochen und die Abkürzung DBDM (sic) sowie die Begriffe *data-informed decision making* und *data use* meist synonym verwendet (Mandinach & Schildkamp, 2021a; Schildkamp & Kuiper, 2010). In der Tendenz wird *data-driven decision making* zunehmend durch die genannten anderen Begriffe ersetzt, um zu verdeutlichen, dass nicht Daten eine Entscheidung (an)treiben sondern die Entscheidung auf Daten basierend bzw. dateninformiert getroffen wird (Brown et al., 2017). Definiert werden kann *data-based decision making* "as systematically analyzing existing data sources within the school, applying outcomes of analyses to innovate teaching, curricula, and school performance, and, implementing (e.g. genuine improvement actions) and evaluating these innovations" (Schildkamp & Kuiper, 2010, S. 482). Das bedeutet, dass unterschiedliche Daten(arten) für Entscheidungen im schulischen Kontext herangezogen werden, um diesbezügliche Prozesse, z.B. den Unterricht betreffend, zu gestalten und zu entwickeln und die Entwicklungsschritte zu überprüfen. Dabei werden Daten als gesammelte und strukturierte Informationen, die einen bestimmten Aspekt von Schule(n) repräsentieren, konzeptualisiert (Lai & Schildkamp, 2013). Es wird also dezidiert von einem weiten Datenbegriff ausgegangen (Mandinach & Schildkamp, 2021a; Schildkamp, 2019), d.h. es geht nicht nur um quantitative formale Daten im Sinne von wissenschaftlichen Kriterien genügenden Ergebnissen aus standardisierten Leistungstests, sondern auch um informelle(re), etwa von Schulen bzw. Lehrkräften selbst erhobene oder vorliegende Daten (z.B. formatives Assessment, Unterrichtsbeobachtungen, Schüler*innenfeedback zum Unterricht, Elternbefragungen, qualitative Interviews sowie Schulstatistiken usw.). Blick man in aktuelle deutschsprachige Überblicksarbeiten (z.B. Thiel et al., 2019; Wurster, 2019) scheint sich der deutsche Begriff der datenbasierten Schul- und Unterrichtsentwicklung hauptsächlich auf die Nutzung von Vergleichsarbeiten, zentralen Abschlussprüfungen, Inspektions- und Evaluationsdaten und den Entwicklungsaspekt zu beziehen. Im englischsprachigen Diskurs wird bei *data-based decision making* neben dem Entwicklungsaspekt auch die datenbasierte Gestaltung von Schule und Unterricht, z.B. durch formatives Assessment, stärker adressiert (z.B. Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Daran angelehnt wird in dieser Arbeit von

datengestützten Entscheidungen bzw. der datenbasierten Gestaltung und Entwicklung von Schule und Unterricht gesprochen.

Mit Blick auf die Digitalisierung im Kontext Schule werden in *data-based decision making* z.B. auch digitale Verhaltensdaten und *big data* grundsätzlich (wenngleich auch relativ unspezifisch) eingeschlossen (z.B. Schildkamp, 2019; Tempelaar et al., 2015). Auch wenn Datafizierung und Digitalisierung zwei unterschiedliche Konzepte und Prozesse sind, erleichtert und verstärkt der Digitalisierungsprozess die Datafizierung, indem z.B. Daten automatisiert erhoben werden, einfacher zur Verfügung stehen und leichter verknüpft werden können (Hartong et al., 2019). Lehrkräfte können so unterschiedliche Datenquellen nutzen oder auch verknüpfen: Mehrebenenperspektivisch betrachtet bezieht sich dies auf der Mesoebene etwa auf eine digitale Infrastruktur (Syring et al., 2022) im Sinne von digitalen Schulmanagement- bzw. Verwaltungssystemen für Stundenpläne, Personalplanung und Fehlzeiten etc. (Hartong et al., 2019). Auf der Ebene des Unterrichts, dessen Vor- und Nachbereitung mit eingeschlossen, sind hier z.B. Daten aus Lernapps, Lernmanagementsystemen, digitalen formativen Assessments oder Feedbacktools zu nennen (Syring et al., 2022). Allerdings bedeutet die verstärkte und erleichterte Verfügbarkeit von Daten weder, dass die Verfügbarkeit mit Konstruktvalidität einhergeht, noch dass die Daten auch für datengestützte Entscheidungen herangezogen werden (Mandinach & Schildkamp, 2021a; Schildkamp, 2019) oder dass sich daraus notwendigerweise (positive) Wirkungen ergeben (vgl. Kapitel 2.2 und 2.3).

Geht es um das Zielkriterium datenbasierter Schul- und Unterrichtsentwicklung bzw. datengestützter Entscheidungen, so wird zwar anhand der Begrifflichkeiten deutlich, dass Schule und Unterricht (weiter)entwickelt, also in irgendeiner Weise verbessert werden sollen. Allerdings bleibt vielfach mehr oder weniger normativ implizit und nicht näher begründet oder diskutiert, dass es hierbei hauptsächlich um (fachliche) Leistungen von Schüler*innen geht (Altrichter et al., 2016). Aktuellere Publikationen nehmen hier verstärkt auch das Wohlbefinden von Schüler*innen (z.B. Mandinach & Schildkamp, 2021) und vor allem *equity* mit auf, d.h. die Adressierung von Bildungsungleichheit und Bildungsgerechtigkeit (z.B. Dodman et al., 2021).

2.2. Prozessmodelle

Während im vorherigen Kapitel einleitende Begriffsklärungen vorgenommen wurden, werden nun zentrale Rahmenmodelle dargestellt und analysiert. Da die vorliegende Arbeit sich auf die Nutzung von Daten von Lehrpersonen fokussiert, stehen Modelle, die auf der Ebene der Lehrpersonen angesiedelt sind, im Mittelpunkt. Daher wird z.B. auf Modelle, die die Ebene der

Einzelschule adressieren, wie etwa das *School Performance Feedback System*-Modell von Visscher & Coe (2003), nicht weiter eingegangen.

Als Ausgangspunkt dient das Modell von Hosenfeld und Groß Ophoff (2007), das eng an ein früheres Modell von Helmke anknüpft (Helmke & Hosenfeld, 2005), und für den Kontext Vergleichsarbeiten ein differenziertes Prozessmodell auf der Ebene von Lehrpersonen darstellt (Altrichter et al., 2016). Es konzeptualisiert verschiedene Prozessschritte (vgl. Abbildung 1), die sequenziell, d.h. nacheinander, durchlaufen werden müssen bzw. können (vgl. lineare Pfeile von links nach rechts): Die Informationen enthaltende Datenrückmeldung muss zuerst rezipiert werden (Rezeption), d.h. die Daten analysiert und verstanden werden, bevor diese im nächsten Schritt reflektiert werden (Reflexion), d.h. z.B. mögliche Erklärungen für die Ergebnisse gefunden werden. Erst dann können Maßnahmen ergriffen und umgesetzt werden (Aktion) und in ihren Wirkungen evaluiert werden (Evaluation). Gerahmt und beeinflusst werden diese Prozessschritte durch individuelle, schulische und (mittelbar) externe Bedingungen.

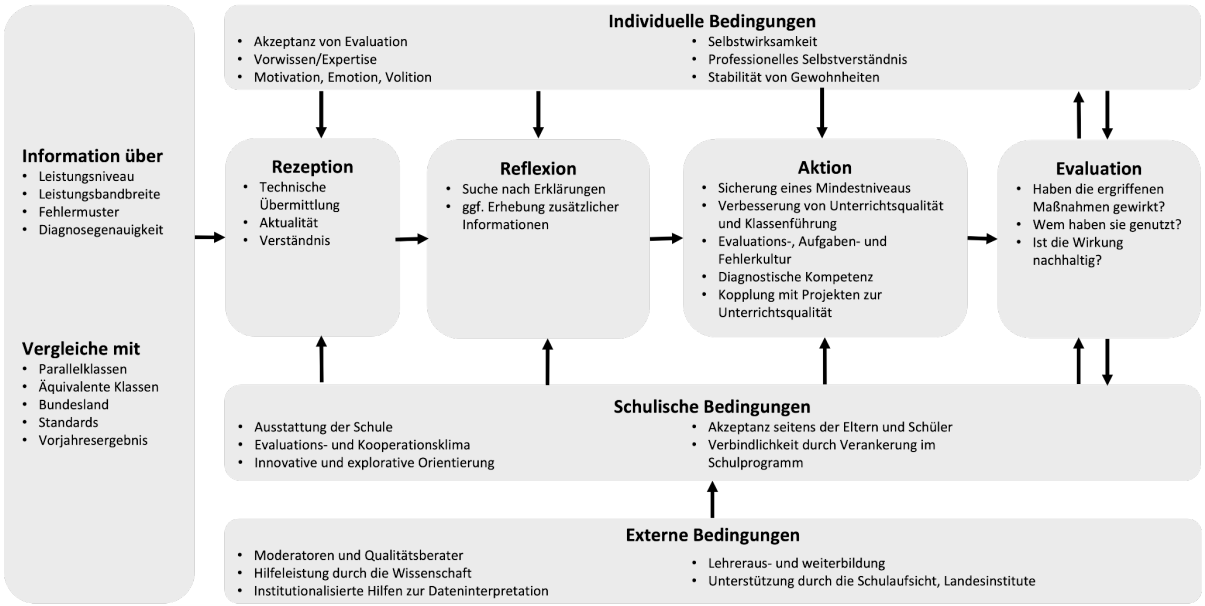


Abbildung 1. Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten nach Hosenfeld und Groß Ophoff (2007)

Da sich dieses Rahmenmodell dezidiert nur auf den Kontext Vergleichsarbeiten bezieht, werden ergänzend weitere Prozessmodelle aus dem internationalen Kontext eingeführt, die datengestützte Entscheidungen allgemeiner und unabhängig von einem bestimmten Verfahren oder einer Datenart konzeptualisieren. Der Fokus liegt dabei auf dem Modell von

Marsh (2012) aufgrund seiner häufiger Rezeption und dem Modell von Schildkamp (2019) aufgrund seiner Aktualität (Bez et al., 2023).

Ähnlich wie beim vorherigen Modell beginnt das Modell von Marsh (2012) mit dem Vorhandensein bzw. dem Zugang zu Daten (1). Im nächsten Schritt werden durch Analyse Informationen aus den Daten gewonnen (2). Entsprechendes Wissen entsteht dann, wenn diese verstanden und mit Expertise und anderen Formen von Wissen, z.B. Kontextinformationen, verbunden werden (3), das dann für Handlungsmaßnahmen angewandt werden kann (4). Deren Wirksamkeit kann daraufhin erfasst werden (5) und der zyklische Prozess beginnt von vorne, wenngleich Feedbackschleifen schon vorher einsetzen können (vgl. gestrichelte Pfeile in Abbildung 2).

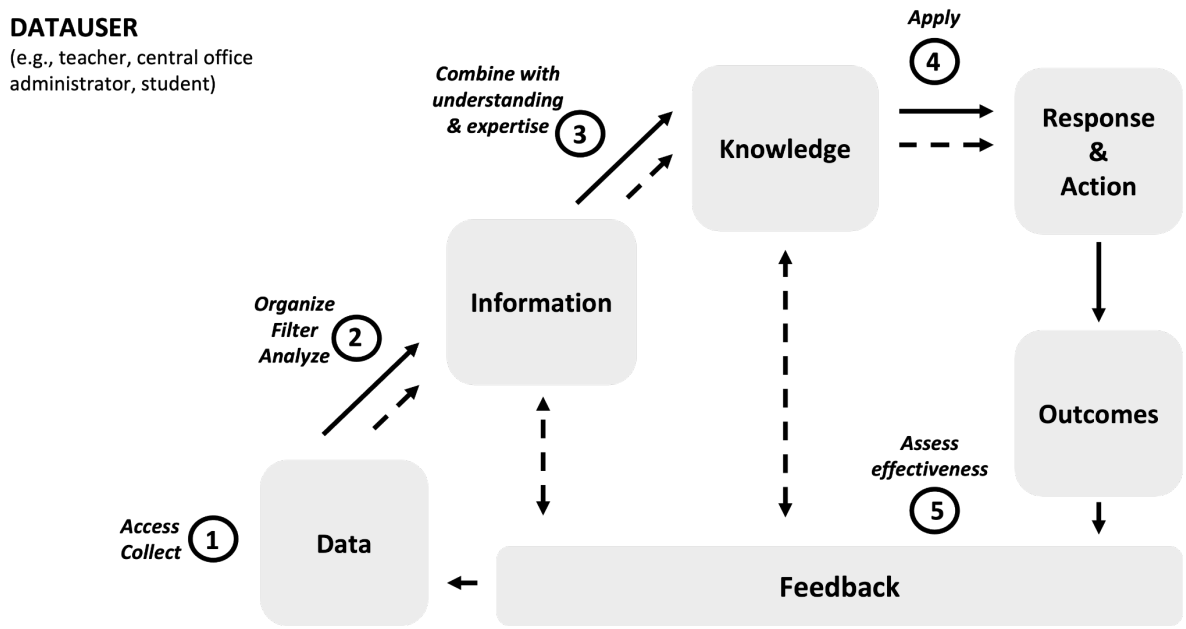


Abbildung 2. Prozessmodell der Datennutzung nach Marsh (2012)

Anders als die beiden bisher dargestellten Modelle beginnt das Prozessmodell von Schildkamp (2019) nicht mit den Daten, sondern mit der normativen Setzung eines Ziels (vgl. Abbildung 3). An diesem orientiert werden dann Daten erhoben. Im Gegensatz zu den beiden anderen vorgestellten Modellen wird *sensemaking*, d.h. die aktive Konstruktion von Bedeutung und Sinn aus den Daten in Verbindung mit Kontextwissen und Expertise durch Analyse und Interpretation der Daten, von Schildkamp zusammenfassend als ein einzelner, gemeinsamer Prozessschritt konzeptualisiert. Allerdings wird im Text die Unterscheidung zwischen Datenanalyse und -interpretation innerhalb von *sensemaking* getroffen (z.B. Schildkamp, 2019, S. 259). Auf *sensemaking* basierend werden dann Schlussfolgerungen für Anschlusshandlungen getroffen und diese evaluiert.

Abbildung 3. Prozessmodell der Datennutzung nach Schildkamp (2019)

In der Zusammenschau der dargestellten Modelle wird deutlich, dass alle Modelle zyklisch-sequenziell aufgebaut sind, d.h. verschiedene Prozessschritte konzeptualisiert werden, die nacheinander durchlaufen werden, bevor ein neuer Zyklus beginnt. Daraus folgt, dass jeder Prozessschritt eine notwendige Voraussetzung für den darauffolgenden Schritt ist. Eine weitere Gemeinsamkeit besteht darin, dass das Vorhandensein von Daten nur eine Grundvoraussetzung ist und sich mögliche Handlungen daraus nicht unmittelbar oder automatisch ergeben, da die Daten rezipiert bzw. analysiert und interpretiert werden müssen. Auch wenn die Modelle sich in den Begrifflichkeiten und den Prozessschritten teilweise unterscheiden, kann aus ihnen konsistent geschlossen werden, dass das Vorhandensein von Daten zwar eine notwendige aber keinesfalls hinreichende Bedingung für den Prozess der datenbasierten Gestaltung und Entwicklung von Schule und Unterricht auf der Ebene der Lehrpersonen ist und aufgrund der notwendigen Teilschritte der Gesamtprozess als komplex gelten kann. Des Weiteren kann aus der sequentiellen Strukturierung aller Modelle abgeleitet werden, dass den ersten Schritten, der Rezeption und Interpretation von Daten, eine besondere Relevanz zukommt, da sie die notwendige Voraussetzung für die weiteren Teilschritte sind.

2.3. Effekte

Nachdem im vorherigen Abschnitt Prozessmodelle zur konzeptuellen Wirkungskette datengestützter Entscheidungen im schulischen Kontext dargestellt wurden, wird nun die empirische Evidenz bezüglich der intendierten Wirkungen in den Blick genommen. Dabei wird auf Forschung eingegangen, die untersucht, welche Effekte datengestützte Entscheidungen

von Lehrpersonen (basierend auf Leistungsdaten) auf die fachlichen Leistungen von Schüler*innen haben, da diese häufig als Zielkriterium gelten (vgl. Kapitel 2.1). Im nächsten Kapitel wird dann differenziert auf förderliche und hinderliche Faktoren eingegangen.

Aus dem deutschsprachigen Kontext liegen hierzu kaum Studien vor, da sich etwa die Begleitforschung zu Vergleichsarbeiten überwiegend aus Befragungsstudien in Form von retrospektiven Selbstauskünften zur Akzeptanz, Nützlichkeit, Informativität und zu abgeleiteten Maßnahmen bezieht, deren eingeschränkte Aussagekraft, auch aufgrund der häufigen Querschnittdesigns, wiederholt problematisiert wird (Altrichter et al., 2016; Dedering, 2011; Hellrung & Hartig, 2013). Hiervon ausgenommen untersuchten Wurster et al. (2017) den Zusammenhang zwischen (selbstberichteter) datenbasierter Unterrichtsentwicklung von Lehrkräften und den Leistungen ihrer Schüler*innen. Hierbei zeigten sich keine signifikanten Zusammenhänge und damit inkonklusive Ergebnisse (Dienes, 2016). Richter et al. (2014) fanden Zusammenhänge zwischen Überzeugungen von Lehrkräften zu verschiedenen Funktionen von Lernstandserhebungen und den Leistungen ihrer Schüler*innen. Dabei zeigten Schüler*innen von Lehrkräften, die Lernstandserhebungen vorrangig eine Entwicklungsfunktion zuschreiben, sowohl in Mathematik als auch im Lesen signifikant bessere Ergebnisse, wobei die Effekte jeweils klein sind. Da sowohl die Studie von Wurster et al. (2017) als auch die Studie von Richter et al. (2014) querschnittliche Designs sind, sind nur Korrelationsanalysen möglich, die keine kausalen Interpretationen erlauben (Rohrer, 2018).

Weitet man den Blick hin zu internationaler Forschung, muss zunächst festgehalten werden, dass Forschungsergebnisse aus *high stakes* Kontexten, z.B. Teile der USA, nicht ohne weiteres auf *low stakes* Kontexte, wie etwa den deutschsprachigen Raum, übertragen werden können (Maag Merki, 2016; Maier & Kuper, 2012). Insbesondere in *high stakes* Kontexten mit einem starken Fokus auf Rechenschaftslegung können nicht intendierte Effekte eintreten, die als problematisch einzuschätzen sind (Mandinach & Schildkamp, 2021a): Unterrichtsbezogen sind hier etwa die Verengung des Curriculums auf bestimmte Inhalte (Au, 2007) bzw. explizites *teaching to the test* (Hamilton et al., 2009b) zu nennen. Dabei können bezogen auf Schüler*innen als schwächer eingeschätzte Lernende von bestimmten (Test-)Situationen ausgeschlossen werden (Ehren & Swanborn, 2012) und Lernende mit niedrigem sozio-ökonomischen Hintergrund marginalisiert werden (Datnow & Park, 2018). Der starke Fokus auf Rechenschaftslegung kann auch dazu führen, dass sich die fachliche Förderung besonders intensiv auf sogenannte *bubble kids* konzentriert, d.h. auf Schüler*innen, die knapp unterhalb einer bestimmten Kompetenzstufe stehen, und als Konsequenz (noch) schwächere Lernende aus dem Blickfeld geraten (Booher-Jennings, 2005).

International liegen einige Interventionsstudien vor, die die gesamte Wirkungskette datenbasierter Entscheidungen von Lehrpersonen (basierend auf Leistungsdaten) auf die fachlichen Leistungen von Schüler*innen in unterschiedlichen Domänen untersuchen, und die in ersten Überblicksarbeiten und Metaanalysen zusammengefasst werden. Im Forschungsüberblick von Visscher (2021) werden mehrere große (quasi-)experimentelle Interventionsstudien zusammengefasst, die in der Primarstufe und Sekundarstufe I über ein bis zwei Schuljahre in den Niederlanden durchgeführt wurden. Dabei konnten in vier von sechs Forschungsprojekten kleine bis moderate positive signifikante Effekte (d = 0.37 bzw. 1-2 *month of growing*) auf die fachliche Leistung von Schüler*innen in interventionsunabhängigen standardisierten Leistungstests in Mathematik, Lesen und Rechtschreibung nachgewiesen werden. Filderman et al. (2018), die in ihrer Metaanalyse datengestützte Leseförderung von Lehrpersonen in Interventionsstudien durchgeführt mit Schüler*innen mit Schwierigkeiten im Leseerwerb in der Primar- und Sekundarstufe zusammenfassen, berichten insgesamt einen substanziellen positiven Effekt von g = 0.24 (95% CI [0.01, 0.46]) im Vergleich zur Kontrollbedingung normalen Unterrichts. Im Vergleich äquivalenter Leseförderungen mit und ohne *data-based decision making* wird der Effekt auf g = 0.27 (95% CI = [0.07, 0.47]) geschätzt, wobei hier nur in der Hälfte der eingegangenen Studien positive signifikante Effekte gefunden wurden und die Anzahl der Studien insgesamt sehr klein ist (Filderman et al., 2018). Eine aktuelle *best evidence* Metaanalyse von Faber et al (2023), die sehr anspruchsvolle Einschlusskriterien für zugrundeliegende Primärstudien anlegt und Effekte von Interventionsstudien bezüglich der Nutzung digitaler Plattformen mit Leistungsdaten von Schüler*innen fokussiert, berichtet insgesamt einen kleinen positiven Effekt von d = 0.12 (95% CI [0.04, 0.19], wobei nicht in allen zugrundeliegenden Primärstudien positive Effekte vorliegen. Ergebnisse bezüglich differenzierterer Analysen, d.h. welche unterschiedlichen Merkmale der Interventionsstudien als besonders effektiv einzuschätzen sind, sind in dieser Metaanalyse inkonklusiv.

Zusammenfassend kann also festgehalten werden, dass die Studienlage, die die Effekte datengestützter Entscheidungen auf die fachlichen Leistungen von Schüler*innen untersucht, insbesondere im deutschsprachigen Raum gering ausgeprägt ist und sich hauptsächlich auf retrospektive Selbstauskünfte, häufig mit querschnittlichen Designs, bezieht. Internationale Studien weisen, insbesondere bei *high stakes* Kontexten, auf potenzielle nicht intendierte problematische Effekte hin. Insgesamt gesehen liegt Evidenz für einen kleinen bis moderaten positiven Effekt von *data-based decision making* von Lehrpersonen auf die fachlichen Leistungen von Schüler*innen in unterschiedlichen Domänen vor. Die Effekte basieren allerdings hauptsächlich auf verschiedenen Interventionsstudien, die sich in ihrem Umfang, ihren Inhalten und Vorgehensweisen deutlich unterscheiden.

2.4. Beeinflussende Faktoren

In diesem Kapitel werden förderliche und hinderliche Faktoren für datengestützte Entscheidungen in den Blick genommen. Die Struktur des Kapitels orientiert sich an Übersichtsarbeiten, die den empirischen Forschungsstand dazu zusammenfassen und zwischen organisationaler Ebene, Daten bzw. Datensystemen und Nutzer*innen unterscheiden (Heitink et al., 2016; Hoogland et al., 2016; Schildkamp et al., 2020).

Auf organisationaler Ebene, womit die Ebene der Einzelschule adressiert wird, werden als relevante Faktoren die Rolle der Schulleitung, eine entsprechende Schulkultur, Kooperation zwischen Personen, eine hinreichende technische Ausstattung bzw. Infrastruktur sowie ausreichend Zeit und Unterstützungssysteme genannt (Heitink et al., 2016; Hoogland et al., 2016; Schildkamp et al., 2014, 2017). Bezogen auf die Daten bzw. die entsprechenden Systeme werden als einflussreiche Faktoren der Zugang, die Verfügbarkeit sowie die Datenqualität aufgeführt (Hoogland et al., 2016; Schildkamp et al., 2014).

Auf Ebene der Nutzer*innen, also der Lehrkräfte, werden das Wissen bzw. die entsprechenden Kompetenzen zur Nutzung von Daten für die Schul- und Unterrichtsentwicklung, im Englischen bezeichnet als *data literacy*, als sehr zentral herausgestellt (Heitink et al., 2016; Hoogland et al., 2016; Schildkamp et al., 2020). Daneben gelten Einstellungen, motivationale Überzeugungen sowie allgemeine Überzeugungen und Einstellungen zu *data-based decision making* als relevant (Heitink et al., 2016; Hoogland et al., 2016; Prenger & Schildkamp, 2018; Schildkamp et al., 2017, 2020).

2.5. Zusammenfassung

In diesem Kapitel wurden zuerst konzeptuelle und begriffliche Grundlagen zur datenbasierten Gestaltung und Entwicklung von Schule und Unterricht als Teil einer sogenannten Evidenzbasierung im Spannungsfeld zwischen Rechenschaftslegung und Entwicklung geklärt (Kapitel 2.1). Dabei wurde herausgearbeitet, dass häufig mehr oder weniger implizit als Zielkriterium (verbesserte) fachliche Leistungen von Schüler*innen gesetzt werden, wobei auch andere Zielkriterien wie das Wohlbefinden von Schüler*innen oder Fragen von Bildungsungleichheit aktuell stärker in den Blick genommen werden. Datengestützten Entscheidungen liegt grundsätzlich ein weiter Datenbegriff zugrunde, der über standardisierte Leistungstests hinausgeht. Durch die zunehmende Datafizierung und Digitalisierung rücken daher vielfältige Daten(arten) verstärkt in den Blick. Durch die zyklisch-sequenzielle Struktur der Prozessmodelle (Kapitel 2.2) wird jedoch schnell deutlich, dass das Vorhandensein von (digitalen) Daten zwar notwendig aber nicht hinreichend ist, da Daten rezipiert, interpretiert und deren Implikationen in Handlungsmaßnahmen transformiert werden müssen, damit diese

Wirkungen sich entfalten und wiederum evaluiert werden können. Des Weiteren impliziert die sequenzielle Struktur, dass der vorherige Prozessschritt eine notwendige Voraussetzung für den folgenden ist, woraus sich eine besondere Relevanz von Datenrezeption und -interpretation ergeben. Denn diese stehen am Anfang des Prozesses und alle folgenden Schritte bauen darauf auf. Bezüglich der empirischen Studien zu den Effekten von *data-based decision making* (Kapitel 2.3) weisen diese vor allem in *high stakes* Kontexten auf nicht-intendierte, problematische Nebeneffekte hin. Überblicksarbeiten basierend auf Interventionsstudien zeigen insgesamt kleine positive Effekte von datengestützten Entscheidungen auf die fachlichen Leistungen von Schüler*innen, wobei nicht alle Primärstudien positive Effekte berichten. Als beeinflussende Faktoren (Kapitel 2.4) gelten auf individueller Ebene neben Überzeugungen, Einstellungen und Motivation von Lehrpersonen vor allem ihre Kompetenzen zur Nutzung von Daten (*data literacy*) als relevant, weshalb letztere im Fokus dieser Arbeit stehen.

3. Data Literacy von Lehrpersonen

In diesem Kapitel zur Kompetenz von Lehrpersonen zur Nutzung von Daten für die Gestaltung und Entwicklung von Schule und Unterricht wird zunächst ein konzeptueller, systematisierender Überblick über unterschiedliche Definitionen und Ansätze gegeben (Kapitel 3.1), bevor anschließend der empirische Forschungsstand skizziert wird (Kapitel 3.2). Da viele Begriffe aus der englischsprachigen Forschung bzw. englischsprachigen Publikationen stammen und nicht immer ohne weiteres äquivalente deutsche Begriffe herangezogen werden können, werden einige englische Begriffe beibehalten und nicht (direkt) übersetzt sondern ggf. im Deutschen paraphrasiert.

3.1. Begriffs- und Konstruktklärung

Um Gemeinsamkeiten und Unterschiede verschiedener Ansätze herauszuarbeiten, wurden im Rahmen dieser Arbeit verschiedene Kategorien erarbeitet, anhand derer eine systematisierende Übersicht erstellt wurde (vgl. Tabelle 1: Synopse zu Ansätzen im Kontext *data literacy*). Dabei wurde der Ansatz von Mandinach und Gummer (2016) als Ausgangspunkt gewählt, da er auf Basis des systematischen Reviews von Beck et al. (2019) als vergleichsweise einschlägig und weit verbreitet gelten kann und sich einige andere Ansätze darauf beziehen (vgl. Tabelle 1).

Ansatz	(Zentrale) Publikation(en)	Bezug zu Lehrpersonen	Datenart/format	Bezug zu Digitalisierung/ Technologie	Gesamter Datennutzungsprozess und Schwerpunkt im Mehrebenensystem	Relationierung zu bestimmten Wissensdomänen
Data literacy for teachers (DLFT)	Mandinach & Gummer (2016)	explizit	breiter Datenbegriff (also nicht nur Leistungsdaten aus standardisierten Verfahren, sondern auch informelle Daten, …)	eher implizit, wird aber nicht ausgeschlossen	gesamter Datennutzungsprozess; Schul- und Unterrichtsentwicklung, d.h. Meso- und Mikroebene; Schwerpunkt auf Unterricht	expliziter Bezug zu Shulman (1987); genaue Relationierung offen
Assessment literacy	Beck et al. (2019); DeLuca et al. (2016)	explizit	alle Formen von Assessment (summativ, formativ, alternative Formen), eher quantitative Daten	eher nicht	Fokus Mikroebene, d.h. Gestaltung und Anpassung von Unterricht. Fokus: fachliches Lernen, z.B. Vorwissen. Evaluation der getroffenen Maßnahmen eher implizit	expliziter Bezug zu Shulman (1987)
Diagnostische Kompetenz/ assessment competence	Herppich et al. (2017)	explizit	eher breit, z.B. auch Daten aus informellen Kontexten wie z.B. Beobachtungen eingeschlossen	eher nicht	Fokus Mikroebene, Schwerpunkt Schüler*innen bzw. deren Merkmale bezogen auf Lernen, traditionell Forschung zu Urteilsakkuratheit von Lehrpersonen; Modelle enden in der Tendenz mit der Diagnose bzw. Entscheidung	Shulman (1987)

Tabelle 1. Synopse zu Ansätzen im Kontext von *data literacy*

Ansatz	(Zentrale) Publikation(en)	Bezug zu Lehrpersonen	Datenart/format	Bezug zu Digitalisierung/ Technologie	Gesamter Datennutzungsprozess und Schwerpunkt im Mehrebenensystem	Relationierung zu bestimmten Wissensdomänen
Statistical literacy	Chick & Pierce (2013); Pierce et al. (2014)	explizit	Daten aus standardisierten summativen Leistungstests	nein	rezipieren und interpretieren von grafisch aufbereiteten Daten; eher Mikroebene, teilweise auch Mesoebene	mathematisch-statistisches Wissen z.B. zu bestimmten Grafiken; Kontextwissen
Graph literacy	Curcio (1987); Friel et al. (2011); Galesic & Garcia-Retamero (2011)	nein	grafisch aufbereitete Daten	nein	rezipieren und interpretieren von grafisch aufbereiteten Daten; keine Einordnung ins Mehrebenensystem möglich	Kontext- bzw. entsprechendes Domänenwissen
Erweiterungen des Ansatzes von Mandinach und Gummer (2016) hinsichtlich Digitalisierung/ Technologie	Cui & Zhang (2022); Kennedy-Clark & Reimann (2022)	explizit	eher breit	explizit	gesamter Datennutzungsprozess; Schwerpunkt eher auf Unterricht	kein neues Modell sondern enger Bezug zu *data literacy for teachers* von Mandinach & Gummer (2016); TPACK (Mishra & Koehler, 2006)

Tabelle 1. Synopse zu Ansätzen im Kontext von *data literacy* (Fortsetzung)

Mandinach und Gummer entwickelten ihr Rahmenmodell *data literacy for teachers* (DLFT) anhand von Expert*inneninterviews, Analysen von Forschungsliteratur sowie Analysen von Dokumenten, in denen US-Bundesstaaten von Lehrpersonen erwartete Kompetenzen aufführen (Mandinach & Gummer, 2016). Auf der Basis dessen definieren sie *data literacy* als

> "the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data (assessment, school climate, behavioral, snapshot, longitudinal, moment-to-moment, etc.) to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn." (Mandinach & Gummer, 2016, S. 367)

Anhand dieser Definition wird ein breiter Datenbegriff deutlich und es wird eine direkte Relationierung zu Domänen professionellen Wissens von Lehrpersonen nach Shulman (1987) vorgenommen. Letztere bleibt jedoch in ihrem genauen Verhältnis etwas unbestimmt, da in der Definition von einer Kombination gesprochen wird, während im weiteren Verlauf des Textes von einer Interaktion die Rede ist (Mandinach & Gummer, 2016, S. 369), die auch in dem Trichtermodell angelegt zu sein scheint (vgl. Abbildung 4).

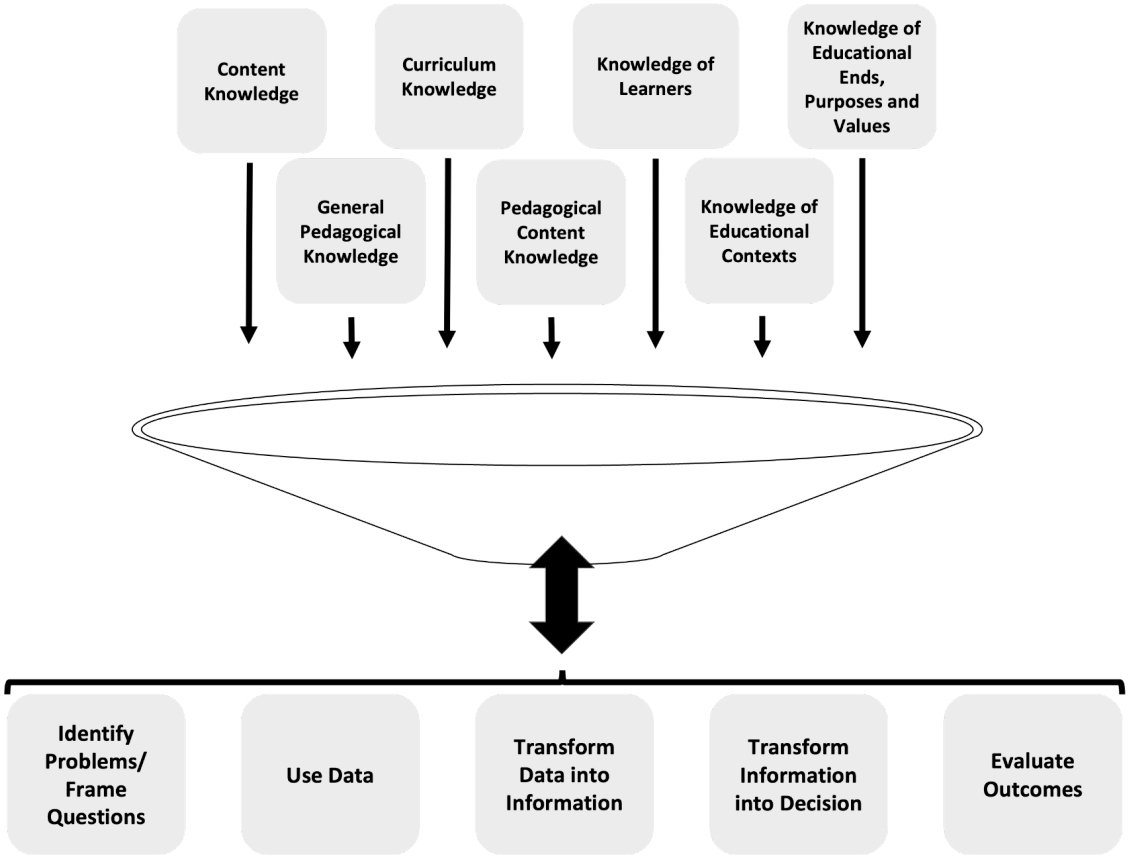


Abbildung 4. *Data literacy for teachers* (DLFT). Rahmenmodell nach Mandinach & Gummer (2016)

Es werden, angelehnt an den Gesamtprozess der Datennutzung, fünf Teilkomponenten von *data literacy* konzeptualisiert, die jeweils noch einmal in zahlreiche Subkomponenten differenziert werden. Diese lassen sich wie folgt zusammenfassen (vgl. dazu auch die Darstellungen in Bez et al., 2023; Wurster et al., 2023):

1) *Identify problems and frame questions*: Zuerst geht es darum, ein bestimmtes Problem zu identifizieren bzw. eine bestimmte Frage zu formulieren, die mithilfe von Daten adressiert werden soll. Dabei können innerhalb der Einzelschule unterschiedliche Ebenen adressiert werden, wie beispielsweise das Curriculum oder der Unterricht.
2) *Use data*: Diese Teilkomponente umfasst die Kompetenzen zur konkreten Datennutzung, d.h. Datenauswahl/-erhebung und Datenanalyse.
3) *Transform data into information*: Diese Teilkomponente umfasst die Transformation der Daten in Information, d.h. die Interpretation der Daten unter Einbeziehung von Kontextinformationen sowie die Suche nach Erklärungen usw.
4) *Transform information into decision*: Diese Teilkomponente zielt auf die Transformation der generierten Informationen in pädagogisch-didaktische Anschlusshandlungen unter Berücksichtigung des konkreten Kontextes sowie deren Umsetzung und Anpassung.
5) *Evaluate outcomes*: Diese Teilkomponente bezieht sich auf die Evaluierung der (intendierten und nicht intendierten) Wirkungen der umgesetzten Maßnahmen vor dem Hintergrund des Eingangsproblems bzw. der Eingangsfrage.

Insgesamt handelt es sich bei dem Rahmenmodell von Mandinach und Gummer (2016) um ein differenziert ausgearbeitetes Modell zu den Kompetenzen von Lehrpersonen für die datenbasierte Gestaltung und Entwicklung von Schule und Unterricht mit dem Schwerpunkt auf Unterricht (Beck et al., 2019; Bez et al., 2023), wobei *use data*, also die Kompetenzen bezüglich der zweiten Komponente der Datennutzung, als Schlüsselkomponente gesehen wird (Mandinach & Gummer, 2016).

Im Kontext von *data literacy* wird immer wieder die Frage nach der Abgrenzung von *assessment literacy* diskutiert (z.B. Beck et al., 2019; Gottlebe et al., 2023; Mandinach & Schildkamp, 2021a). Beck et al. (2019) sehen *assessment literacy* als das engere Konstrukt, da es sich rein auf unterschiedliche Formen von Assessments bezieht, und konzeptualisieren *assessment literacy* daher als einen Teil von *data literacy*. Dieser Argumentation schließt sich die vorliegende Arbeit an. *Assessment literacy* bezieht sich außerdem hauptsächlich auf das fachliche Lernen, wenn es etwa darum geht, das Vorwissen von Schüler*innen zu diagnostizieren, um Unterricht entsprechend zu gestalten und Anpassungen vorzunehmen

(z.B. Gottheiner & Siegel, 2012) und hat einen expliziten Bezug zu messtheoretischen Gütekriterien wie Objektivität, Reliabilität und Konstruktvalidität (Beck et al., 2019; DeLuca et al., 2016).

Ein weiteres benachbartes Konstrukt ist *assessment competence*, was im deutschsprachigen Kontext dem entspricht, was als diagnostische Kompetenz von Lehrpersonen bezeichnet wird (Gottlebe et al., 2023; Herppich et al., 2018). Allerdings werden *assessment literacy* und *assessment competence* im Englischen durchaus synonym verwendet (vgl. z.B. DeLuca et al., 2016; Gottlebe et al., 2023; Pastore, 2023), was bei der Abgrenzung zu einer eher fachlich-fachdidaktisch konzeptualisierten *assessment literacy* (vgl. vorheriger Absatz) beachtet werden muss. Daher wird im Folgenden der deutsche Begriff der diagnostischen Kompetenz von Lehrpersonen verwendet. Bei empirischen Studien zur (Qualität der) diagnostischen Kompetenz von Lehrpersonen geht es häufig um die Urteilsakkuratheit, also wie exakt Lehrkräfte die Leistungen von Schüler*innen beurteilen können, z.B. bei der Bildung einer Rangfolge innerhalb einer Lerngruppe (Schrader & Praetorius, 2018). Auch hier gibt es einen Bezug zu Daten, um bestimmte Informationen, auf deren Basis eine Einschätzung bezüglich eines bestimmten Ziels erfolgt, zu nutzen. Hierbei ist der Datenbegriff ebenso weiter gefasst. Allerdings liegt der Fokus auf der Ebene der Schüler*innen, genauer, auf Merkmalen, die für das Lernen von Schüler*innen relevant sind (Herppich et al., 2018). Das bedeutet, dass auch hier ein klarer Bezug zu Shulman (1987) gegeben ist. Während beide Ansätze prozesshaft angelegt sind, unterscheiden sie sich im Hinblick auf ihren Endpunkt: Während beim Diagnostizieren tendenziell die Diagnose und die sich anschließende Entscheidung den Abschluss bilden, wird bei *data literacy* die Evaluation der umgesetzten Handlungen dezidiert mit eingeschlossen.

Ein weiterer Ansatz im Kontext von *data literacy* ist *statistical literacy* (Chick & Pierce, 2013; Pierce et al., 2014). Dabei geht es nicht, wie der Begriff erst einmal nahelegen würde, um eine allgemeine statistische Kompetenz, sondern um die (benötigten) Kompetenzen von Lehrpersonen, um grafisch aufbereitete Ergebnisse aus standardisierten (summativen) Leistungstests zu rezipieren und zu interpretieren. Dabei wird vorrangig eine Relationierung zu mathematisch-statistischem Wissen, z.B. zu Boxplots, sowie Kontextwissen vorgenommen. *Statistical literacy* wird hierbei als hierarchisch aufgebaut angenommen. Hierbei beziehen sich die Autor*innen auf allgemeinere Ansätze zu *graph literacy* (s. unten) wie z.B. Curcio (1987). Diese wird vergleichsweise eng an bestimmten grafischen Darstellungen, die sich (zum damaligen Zeitpunkt) in entsprechenden Rückmeldungen an Schulen und Lehrkräfte in Australien fanden, ausgearbeitet (Chick & Pierce, 2013). Dabei wird das Ablesen direkt gegebener Werte bzw. Informationen (niedriges Niveau) von dem

Vergleichen von Werten (mittleres Niveau) und der Analyse bzw. Verstehen der gesamten Daten (höchstes Niveau) unterschieden.

Auch wenn der Ansatz der *graph literacy* (Curcio, 1987; Friel et al., 2001; Galesic & Garcia-Retamero, 2011) ursprünglich keinen expliziten Bezug zu den Kompetenzen von Lehrpersonen zur Nutzung von Daten für die Gestaltung und Entwicklung von Schule und Unterricht hat, wird dieser kurz skizziert: Zum einen, weil er der Konzeptualisierung von *statistical literacy* zugrunde liegt (s. oben), und zum anderen, weil auf ihn in empirischen Studien zur Rezeption und Interpretation von Daten von Lehrpersonen Bezug genommen wird (z.B. im Kontext Lernverlaufsdiagnostik: Espin et al., 2017; Zeuch et al., 2017; siehe auch Kapitel 3.2.1). Hier werden überwiegend drei Niveaustufen angenommen: *reading the data*, die unterste Stufe, umfasst das direkte Ablesen einzelner Werte oder Datenpunkte, z.B. einen direkt in der Grafik dargestellten Mittelwert. Von *reading between data*, der mittleren Stufe, wird ausgegangen, wenn Relationierungen zwischen einzelnen Datenpunkten vorgenommen werden, z.B. größer-kleiner-Relationen oder die Bildung von Gruppen. *Reading beyond data*, die höchste Stufe, umfasst die Erfassung der Gesamtinformation der Grafik sowie die Fähigkeit, Schlussfolgerungen zu ziehen, z.B. bei einer Grafik, die einen Trend in den Daten darstellt. Bei den höheren Stufen wird somit mehr Kontext- bzw. entsprechendes Domänenwissen miteinbezogen und von einzelnen Datenpunkten immer weiter abstrahiert.

Abschließend werden nun zusammengefasst Ansätze behandelt, die an sich keine neuen Ansätze von *data literacy* sind, sondern das Rahmenmodell von Mandinach und Gummer (2016) auf das TPACK-Modell (Mishra & Koehler, 2006) beziehen. Cui und Zhang (2022) führten dazu Fokusgruppeninterviews mit Lehrkräften zur Relationierung von TPACK und *data literacy*. Dabei erscheint die empirische Vorgehensweise bei der Entwicklung des Modells in Form eines Venn-Diagramms diskussionswürdig, da viele Lehrpersonen mit den Modellen von TPACK und *data literacy* nicht vertraut waren und zudem konfirmatorische Schlüsse auf der Basis eines explorativen Forschungsdesigns gezogen werden. Kennedy-Clark und Reimann (2022) schlagen aus einer Professionalisierungsperspektive für ein gemeinsames Modell von *data literacy* und TPACK die Metapher eines Rhizoms vor: Damit wird von der Struktur des Venn-Diagramms Abstand genommen und eine netzwerkartige Strukturierung in Form eines Wurzelknotens vorgeschlagen. Allerdings scheint dabei unklar, wie dieser Ansatz als Basis von Operationalisierungen in empirischen Studien genutzt werden könnte.

Zusammenfassend liegen unterschiedliche Konzeptualisierungen von *data literacy* von Lehrpersonen vor. Diese lassen sich hinsichtlich der Bezugnahme auf Daten und Digitalisierung/Technologie, unterschiedlicher Schwerpunkte im Mehrebenensystem sowie im

Datennutzungszyklus und der Relationierung zu bestimmten Wissensdomänen unterscheiden. Mit Blick auf die systematisierende und vergleichende Synopse (siehe Tabelle 1) kann dabei das Rahmenmodell von Mandinach und Gummer (2016) als vergleichsweise breit und umfangreich gelten und bildet daher einen zentralen Bezugspunkt dieser Arbeit.

3.2. Empirischer Forschungsstand

Nachdem im vorherigen Kapitel ein Überblick über konzeptionelle Ansätze von *data literacy* gegeben wurde, wird nun der empirische Forschungsstand zu den Kompetenzen von Lehrpersonen im Umgang mit Daten dargestellt. Dabei wird zunächst allgemeiner auf die Ausprägung bei Lehrpersonen eingegangen, woraufhin auf Studien zur Förderung von *data literacy* bei angehenden Lehrpersonen eingegangen wird.

3.2.1. Ausprägung von *data literacy* bei Lehrpersonen

Um einen strukturierten Überblick über die Ausprägung von *data literacy* von Lehrpersonen anhand der vorliegenden Studien geben zu können, wurde eine Gliederung über verschiedene forschungsmethodische Zugänge gewählt. So können die Einzelergebnisse gebündelt vor dem Hintergrund der Stärken und Limitationen der jeweiligen Zugänge zur Erfassung von *data literacy* von Lehrpersonen eingeordnet und diskutiert werden.

3.2.1.1. Erkenntnisse basierend auf retrospektiven Selbstauskünften in quantitativen Fragebogenstudien

Nach der Einführung von Lernstandserhebungen wurden in Deutschland Lehrkräfte in quantitativen Fragebogenstudien in Form von retrospektiven Selbstauskünften u.a. zur eingeschätzten Verständlichkeit der Ergebnisse und zu ihrem Umgang mit den Ergebnissen befragt (Altrichter et al., 2016), deren Ergebnisse nun zusammenfassend skizziert werden. Die Rückmeldungen werden insgesamt als verständlich eingeschätzt und es findet eine mehr oder weniger intensive individuelle Auseinandersetzung damit statt (Groß Ophoff, 2013b; Koch, 2011). Allerdings werden insgesamt eher selten Veränderungen im Unterricht vorgenommen, also konkrete Maßnahmen auf Basis der Ergebnisse ergriffen (Altrichter et al., 2016): Wenn, dann beziehen sich diese eher auf kleinere Anpassungen im Unterricht, wie z.B. die Weiternutzung der Aufgaben oder Wiederholung von Inhalten, und weniger auf weitreichende strukturelle Veränderungen (Dedering, 2011). Allerdings müssen bei Studien, die auf retrospektiven Selbstauskünften basieren, immer Limitationen mit Blick auf tatsächliches Verhalten beachtet werden (Döring et al., 2016). Pointiert formulieren dies Altrichter et al. hinsichtlich der dargestellten Studienlage aus dem deutschsprachigen Raum folgendermaßen: “Über [...] die *tatsächlichen Handlungen* aller Beteiligten und darüber, welche *Auswirkungen* eventuelle Maßnahmen auf Lernen und Lernergebnisse hatten, wissen

wir aus deutschsprachigen Untersuchungen praktisch nichts. Überdies dürfte sich der Aussagewert von Fragebogendaten in einem sensiblen Bereich wie der Evaluation und Entwicklung von Unterricht in engen Grenzen halten" (Altrichter et al., 2016, S. 249; Hervorhebungen im Original). Dazu korrespondierend korrelierte in der Studie von Schliesing (2017) im Kontext von Vergleichsarbeiten die Verständlichkeit der Rückmeldungen erfasst mit Selbstauskünften mit dem tatsächlichen Verständnis erfasst in einem Test nur gering (r = .24).

3.2.1.2. Erkenntnisse basierend auf Tests

Zur Frage der Ausprägung von *data literacy* ist es sinnvoll, Interventionsstudien, die die Förderung von *data literacy* bei Lehrpersonen adressieren, heranzuziehen, weil sie *data literacy* mithilfe von Tests erfassen. Koch (2013) etwa entwickelte einen Test zur Datenkompetenz von Lehrkräften mit dem Schwerpunkt Datenanalyse und -interpretation bei Vergleichsarbeiten. Dabei konnten im Prätest nur 20% der Lehrkräfte Aufgaben des höchsten Niveaus (das ungefähr *reading beyond data* entspricht, siehe Kapitel 3.1) lösen und 30% konnten maximal sehr einfache bis einfache Aufgaben (entspricht ungefähr *reading the data*, siehe Kapitel 3.1) lösen. Im Posttest schnitten die Lehrpersonen allerdings deutlich besser ab (d = 1.08, p < .000). Aus der internationalen Forschung können weitere Interventionsstudien herangezogen werden, in denen die Kompetenzen von Lehrpersonen im Umgang mit Daten in Tests erfasst wurden. Ebbeler et al. (2017) erfassten im Rahmen einer quasiexperimentellen Interventionsstudie die *data literacy* von Lehrpersonen in einem Test, der Items zu allen Teilkompetenzen enthielt, aber relativ nah am Interventionsinhalt konzipiert wurde. Dabei wurde im Prätest im Mittel ein Wert knapp unterhalb der Hälfte des Maximalwerts von 22 Punkten erreicht (Prätest: MW = 9.4, SD = 3.16; Posttest: MW = 10.4; SD = 3.05; d = .32, p = .005). Im Rückgriff auf den Ansatz von Mandinach und Gummer (2016) entwickelten Kippers et al. (2018) den Test von Ebbeler et al. (2017) weiter und setzten ihn wiederum in einer Interventionsstudie mit Lehrkräften in den Niederlanden ein. Ähnlich wurde auch hier im Prätest im Mittel ein Wert unterhalb der Hälfte des Maximalwerts erreicht (Prätest: Max = 25, MW = 9.3, SD = 2.66; Posttest: MW = 11.2, SD = 3.03; d = 0.71; p = .004). Dabei schnitten die Lehrpersonen bei den Subkompetenzen Zielsetzung/Frageformulierung und Datenanalyse am schlechtesten ab. van der Kleij und Eggen (2013) untersuchten, wiederum in den Niederlanden, wie gut Lehrpersonen im Vergleich zu anderen Akteursgruppen eine Plattform mit Lernstandserhebungen und -verläufen ihrer Schüler*innen analysieren und interpretieren konnten. Auch hier erreichten die Lehrpersonen im Mittel ungefähr die Hälfte des Maximalwerts des Tests (Max = 34, MW = 17.8, SD = 8.7). Dabei korrelierte die selbsteingeschätzte Kompetenz eher gering mit der im Test erfassten (r = .25).

Zusammenfassend deuten Studien, in denen *data literacy* mit Tests erfasst wurde, eher auf gering bis moderat ausgeprägte Kompetenzen bei Lehrpersonen hin. Dies muss allerdings vor dem Hintergrund methodischer Limitationen eingeordnet werden: Zunächst muss einschränkend formuliert werden, dass viele Tests im Kontext von Interventionen entwickelt wurden und daher inhaltlich recht nahe am Interventionsinhalt oder einem bestimmten Verfahren wie z.B. Vergleichsarbeiten sind und damit potenziell bezüglich ihrer Konstruktvalidität Einschränkungen aufweisen. Des Weiteren basieren die meisten Tests auf einfachen Summenscores und wurden unterschiedlich streng auf ihre psychometrische Güte hin geprüft. Außerdem kann der Schwierigkeitsgrad der einzelnen Tests eher schwer eingeschätzt werden und alle Studien basieren auf Gelegenheitsstichproben. Abschließend weisen Ebbeler et al. (2017) zurecht darauf hin, dass man auf der Basis von (Veränderungen von) *data literacy* gemessen in Tests im Rahmen von Interventionsstudien nicht direkt auf die Performanz von Lehrkräften in der alltäglichen Praxis schließen kann und betonen diesbezüglichen Forschungsbedarf.

3.2.1.3. Erkenntnisse basierend auf Interviews und lautem Denken

Ergänzend zu Fragebogenstudien und Tests wurden Studien mit leitfadengestützten Interviews sowie Studien mit lautem Denken durchgeführt, die die Kompetenzen von Lehrpersonen im Umgang mit Daten adressieren. Means et al. (2011) führten in den USA in Einzel- und Gruppensettings Vignetteninterviews durch, um die Kompetenzen von Lehrpersonen im Umgang mit Assessmentdaten für die Unterrichtsentwicklung zu explorieren. Dabei zeigte sich, dass Lehrpersonen generell direkt dargestellte Werte und Informationen korrekt entnehmen konnten und unterrichtliche Anpassungen diskutierten, während komplexere Elaborationen der Datenanalyse und -interpretation und die Formulierung zum Assessmentsystem passender Problemstellungen für die Unterrichtsentwicklung Schwierigkeiten bereiteten. van der Kleij und Eggen (2013) führten Fokusgruppeninterviews zur Datenanalyse und -interpretation der Rückmeldungen eines Onlinesystems von Lernstandserhebungen in den Niederlanden. Dabei riefen einige Darstellungselemente und statistische Entitäten wie Benchmarks und Konfidenzintervalle Verwirrung und falsche Interpretationen hervor. Ähnlich konnten Goffin et al. (2023) verschiedene Fehlkonzepte und Fehler bezüglich der Analyse und Interpretation von authentischen Rückmeldungen nach landesweiten standardisierten Leistungstests bei belgischen Lehrkräften herausarbeiten, die sie als mangelnde Passung zwischen Elementen der Rückmeldungen und zugrundeliegenden Konzepten und Kompetenzen von Lehrpersonen interpretieren. Auch Gutwirth et al. (2021) führten leitfadengestützte Interviews mit belgischen Mathematiklehrkräften (der Sekundarstufe) zu Rückmeldungen landesweiter standardisierter Leistungstests mit dem Ergebnis, dass auch Mathematiklehrkräfte, denen höhere Expertise im Umgang mit Daten

zugesprochen werden könnte, die Daten häufig eher unstrukturiert und oberflächlich analysierten und sie teilweise fehlerhaft interpretierten.

Neben den aufgeführten Interviewstudien, bei denen der Umgang mit standardisierten Testergebnissen im Vordergrund steht, gibt es Studien aus dem Bereich Lernverlaufsdiagnostik, in der mithilfe von lautem Denken untersucht wurde, wie Lehrkräfte Lernverlaufsdiagnostikgrafiken rezipieren, interpretieren und Schlussfolgerungen für den Unterricht ziehen. Dabei wurde gleichzeitiges und nicht retrospektives lautes Denken angewandt, d.h. Lehrpersonen gebeten, während einer Aufgabe ihre Gedanken und Überlegungen laut zu äußern, und nicht rückblickend nach der Aufgabe (Leighton, 2017). Dabei zeigte sich in einer Studie von van den Bosch et al. (2017), dass Lehrpersonen wenig Probleme damit hatten, direkt gegebene Werte abzulesen, was der niedrigsten *graph literacy*-Stufe *reading the data* entspricht (Friel et al., 2001). Aber wie beim Umgang mit standardisierten Leistungstests zeigten sich Schwierigkeiten, wenn es darum ging, die Daten zu interpretieren und Schlussfolgerungen für unterrichtliche Anschlusshandlungen zu formulieren. Dabei zeigten sich keine signifikanten positiven Korrelationen zwischen der allgemeinen *graph literacy* erfasst in einem Test und der Performanz der Lehrkräfte erfasst mithilfe des lauten Denkens ($\tau < 0.09$), wobei der Test allerdings Deckeneffekte aufwies (van den Bosch et al., 2017). Eine ähnliche Studie führten Espin et al. (2017) durch und untersuchten die *think aloud*-Protokolle von Sonderschullehrkräften hinsichtlich der Komplexität, Akkuratheit und Kohärenz mithilfe von Expertenratings. Auch hier gab es Lehrkräfte, die die höchste Stufe *reading beyond data* erreichten, während andere auch Schwierigkeiten beim Ablesen direkter Werte und bei der Formulierung kohärenter Schlussfolgerungen zeigten. Ein ähnliches Ergebnis zeigte sich in den *think aloud*-Protokollen in der Studie von Zeuch et al. (2017) mit deutschen Lehrkräften, wobei die Lehrkräfte gleichzeitig in einem lernverlaufsdiagnostikspezifischen Test relativ gut abschnitten (*Max* = 134, *MW* = 104.2, *SD* = 18.0).

Zusammenfassend liegen bisher einzelne (explorative) Studien vor, die mithilfe unterschiedlicher Interviewverfahren und inhaltsanalytischer Auswertung die *data literacy* von Lehrpersonen adressieren. Dabei liegt bei den meisten Studien der Schwerpunkt auf der allgemeinen Datenrezeption und -interpretation bei mehr oder weniger authentischen Ergebnisrückmeldungen. Nur bei Goffin et al. (2023) und van den Bosch (2017) lagen teilweise Ergebnisgrafiken der eigenen Klassen der teilnehmenden Lehrkräfte zugrunde. Insgesamt zeigt ein (tendenziell eher kleinerer) Teil der Lehrkräfte Elaborationen auf der höchsten *graph literacy*-Stufe und kann damit die Gesamtaussage der Daten erfassen, während dieser Schritt für den Großteil der Lehrkräfte in den vorliegenden Studien mit

Schwierigkeiten verbunden ist. Dabei bestehen Einschränkungen zum einen mit Blick auf die ökologische Validität, da die zugrundeliegenden Ergebnisse keine breite Datengrundlage abbilden und selten aus den eigenen Klassen stammen. Zum anderen bestehen Einschränkungen aufgrund der kleinen Gelegenheitsstichproben, die generalisierende Aussagen auf der Basis von (qualitativeren und quantifiziererenden) Inhaltsanalysen stark limitieren. Weiterhin erlaubt die Erhebungsmethode des gleichzeitigen lauten Denkens grundsätzlich, die erhobenen Daten aus einer Prozessperspektive zu analysieren, weil bei bestimmten Stimuli angenommen wird, dass die Gedankensequenzen durch die Verbalisierung nicht betroffen sind (Ericsson & Simon, 1998). Allerdings werden in den dargestellten Studien überwiegend Häufigkeiten auf der Basis inhaltsanalytischer Kodierungen ausgewertet und damit eine Stärke der Methode, nämlich eine Prozessperspektive einzunehmen, kaum genutzt.

3.2.2. Förderung von *data literacy* bei angehenden Lehrpersonen

Wie in den vorangegangenen Teilkapiteln herausgearbeitet wurde, gilt *data literacy* von Lehrkräften als relevant für das Gelingen datenbasierter Entscheidungen im schulischen Kontext. Gleichzeitig deuten Studien, die Aussagen zur Ausprägung von *data literacy* bei Lehrpersonen erlauben, insgesamt darauf hin, dass die entsprechenden Kompetenzen bei einem Großteil von Lehrpersonen eher gering bis moderat ausgeprägt sind. Vor diesem Hintergrund wird auf die Relevanz der Anbahnung entsprechender Kompetenzen bereits in der Ausbildung von Lehrpersonen hingewiesen (z.B. Beck & Nunnaley, 2021; Mandinach & Gummer, 2016). Die Dringlichkeit wird dadurch verstärkt, dass in den meisten Curricula die entsprechenden Kompetenzen bisher nicht adressiert werden (Merk et al., 2020).

Korrespondierend liegen erste Interventionsstudien dazu vor, die untersuchen, wie *data literacy* wirksam bei Lehramtsstudierenden gefördert werden kann. Dabei ist die Zahl dieser Studien (noch) relativ klein (Reeves & Honig, 2015). Dies kann man auch daran sehen, dass in einer aktuellen Metaanalyse von Filderman et al. (2022) zur Förderung von *data literacy* bei Lehrpersonen nur 7 von 33 zugrundeliegenden Studien Lehramtsstudierende adressieren. In dieser Metaanalyse gehen noch weniger Studien (zumindest teilweise) über die Erfassung der adressierten Kompetenzen mithilfe von Selbsteinschätzungen hinaus. Reeves und Honig (2015) entwickelten und evaluierten eine *data literacy*-Intervention für Lehramtsstudierende in einem Prä-post-Design, die sich inhaltlich auf summative Assessments für die Unterrichtsentwicklung bezog. Dabei berichten sie einen moderaten positiven signifikanten Effekt ($d = 0.6$, $p < .001$) der Intervention auf die Kompetenzen der teilnehmenden Studierenden (Reeves & Honig, 2015). Eine andere Studie aus dem amerikanischen Kontext mit dem Fokus auf schulexternen standardisierten Leistungsdaten, ebenfalls in einem Prä-

post-Design, konnte die Kompetenzen von Lehramtsstudierenden in der Größenordnung eines kleinen positiven Effekts ($d = 0.34$, $p < .05$) signifikant steigern (Reeves & Chiang, 2017). Aus dem deutschsprachigen Kontext liegen weitere Studien vor, die jeweils spezifische Teilkompetenzen in ihren Interventionsstudien fokussieren: Merk et al. (2020) entwickelten und überprüften in einem *randomized controlled trial* (RCT) in einem Wartekontrollgruppendesign eine Intervention zur Datenanalyse und -interpretation für Lehramtsstudierende und berichten einen moderat bis großen positiven signifikanten Effekt ($\beta = 0.56$, $p < .001$). Jungjohann et al. (2022) entwickelten und überprüften eine kurze videobasierte Onlineintervention (in Form eines RCT) zur *graph literacy* bei Lernverlaufsdiagnostik und berichten einen moderaten positiven signifikanten Effekt (*Cohens´ f* = 0.39, $p < .001$). Einschränkend bestand hier die Stichprobe aus Lehrpersonen (gut 40%) und Lehramtsstudierenden (knapp 60%) und es wurden keine Analysen getrennt für die beiden Gruppen durchgeführt. Insgesamt liegen also erste Studien zur Förderung von *data literacy* von Lehramtsstudierenden vor, die signifikante positive Effekte berichten. Allerdings adressiert der überwiegende Teil der Studien nur Teilkompetenzen bestimmter Konzepte von *data literacy* bzw. bezieht sich nur auf die Rezeption und Interpretation von Leistungsdaten von Schüler*innen.

3.3. Zusammenfassung

In diesem Kapitel wurde zunächst ein systematisierender Überblick über verschiedene Definitionen und Ansätze von *data literacy* von Lehrpersonen gegeben (Kapitel 3.1). Diese lassen sich anhand der Kriterien des expliziten Bezugs zu Lehrpersonen, der Datenart/form, des Bezugs zu Digitalisierung/Technologie sowie der Einordnung im gesamten Datennutzungszyklus, dem Mehrebenensystem sowie der Relationierung zu bestimmten Wissensdomänen unterscheiden. Dabei kann der Ansatz von Mandinach und Gummer (2016) als vergleichsweise umfassend und differenziert gelten. Anschließend wurde der empirische Forschungsstand zur Ausprägung von *data literacy* anhand verschiedener methodischer Zugänge (retrospektive quantitative Befragungen, Tests in Interventionsstudien, verschiedene Formen von Interviews) dargestellt (Kapitel 3.2). Insgesamt deuten die bisherigen Studien darauf hin, dass die entsprechenden Kompetenzen bei Lehrpersonen eher gering bis moderat ausgeprägt sind. Lehrpersonen scheinen insbesondere Schwierigkeiten bei der Datenrezeption und -interpretation zu haben, was etwa die Gesamterfassung von in Grafiken visualisierten Daten anbelangt. Allerdings sind aufgrund der dargestellten Limitationen nur eingeschränkt Rückschlüsse auf die tatsächliche Performanz der Lehrpersonen in der Praxis im Sinne ihres alltäglichen Umgangs mit Daten für die datenbasierte Gestaltung und Entwicklung von Schule und Unterricht möglich. Im letzten Teilkapitel (Kapitel 3.3) wurde auf die Förderung von *data literacy* bei angehenden Lehrpersonen eingegangen. Deren Relevanz

lässt sich durch die eher gering ausgeprägten Kompetenzen bei Lehrpersonen sowie der bisher kaum vorhandenen curricularen Verankerung in der ersten Phase der Lehrpersonenbildung begründen. Es liegen erste Interventionsstudien dazu vor, die positive signifikante Effekte zeigen. Allerdings beziehen sie sich vorrangig auf Leistungsdaten und adressieren jeweils nur Teilkompetenzen von *data literacy*.

4. Forschungsinteresse

In diesem Kapitel wird das Forschungsinteresse der Arbeit, das sich auf die Datenrezeption und -interpretation von Lehrpersonen bei datengestützten Entscheidungen bezieht, expliziert. Dazu werden zunächst die Forschungsdesiderate, die in den vorherigen Kapiteln bereits angeklungen sind, gebündelt. Daraufhin werden die konkreten Forschungsfragen dieser Arbeit formuliert und ihre Adressierung in den verschiedenen durchgeführten Forschungsarbeiten (Kapitel 5-8) im Überblick skizziert.

4.1. Forschungsdesiderate

Schule und Unterricht datenbasiert zu gestalten und weiterzuentwickeln, gilt als eine Aufgabe von Lehrpersonen. Dies wird als zyklisch-sequenzieller Prozess konzeptualisiert, in dem der vorherige Schritt die notwendige Voraussetzung für den kommenden ist (vgl. Kapitel 2.2). Dabei reicht die alleinige Verfügbarkeit von Daten für die angenommene Wirkungskette bis hin zu den intendierten Effekten im Sinne verbesserter Leistungen von Schüler*innen nicht aus (Schildkamp, 2019). Dies gewinnt durch die zunehmende Digitalisierung und die damit einhergehende verstärkte und erleichterte Verfügbarkeit von Daten an Relevanz. Überblicksarbeiten, die bisherige Studien zu beeinflussenden Faktoren von *data-based decision making* zusammenfassen, nennen die *data literacy* von Lehrpersonen, also die Kompetenzen zur Nutzung von Daten, als (einen) zentralen Faktor (z.B. Heitink et al., 2016; Hoogland et al., 2016). Daher steht sie im Zentrum dieser Arbeit.

Im Prozess datengestützter Entscheidungen kommen der Datenrezeption und -interpretation besondere Bedeutung zu: Diese stehen am Beginn des Prozesses, weshalb alle folgenden Schritte, wie etwa die Konstruktion von konkreten Anschlusshandlungen und deren Umsetzung, darauf aufbauen. Gleichzeitig gelten Studien mit dem Fokus auf den Alltag von Lehrkräften als Desiderat (Mandinach & Schildkamp, 2021a): Quantitative Fragebogenstudien in Form retrospektiver Selbstauskünfte, z.B. im Rahmen der Begleitforschung zu Vergleichsarbeiten, erlauben nur wenige Rückschlüsse auf die tatsächlichen Handlungen und Kognitionen von Lehrpersonen im schulischen Alltag (Döring et al., 2016), sind also bezüglich

ihrer ökologischen Validität und der Prozessperspektive eingeschränkt. Zudem korreliert die eingeschätzte Verständlichkeit der Rückmeldungen nur gering mit dem tatsächlichen Verständnis (Schliesing, 2017). Tests im Rahmen von Interventionsstudien zur Förderung von *data literacy* deuten insgesamt auf eher gering bis moderat ausgeprägte Kompetenzen von Lehrpersonen hin, wobei diese potenziell aufgrund der Nähe zu bestimmten Interventionsinhalten in ihrer Konstruktvalidität eingeschränkt sind und Kompetenzen gemessen in Testscores schwer direkte Rückschlüsse auf die Prozesse und die Performanz von Lehrkräften in ihrem Alltag ermöglichen (vgl. Kapitel 3.2.1.2). Durchgeführte Interviewstudien mit inhaltsanalytischer Auswertung zeigen auch tendenziell Schwierigkeiten von Lehrpersonen bei der Datenrezeption und -interpretation (z.B. Goffin et al., 2023; Zeuch et al., 2017). Hier ist die ökologische Validität, vor allem bei Studien mit lautem Denken, höher einzuschätzen, wobei Studien, die mit eigenen Daten der Lehrperson durchgeführt wurden, rar sind (vgl. Kapitel 3.2.1.3). Zudem ermöglicht die Methode des lauten Denkens grundsätzlich eine Prozessperspektive (Ericsson & Simon, 1998), jedoch kaum bei einer bisher dominierenden inhaltsanalytischen Kodierung mit anschließender Auswertung von Häufigkeiten (vgl. Kapitel 3.2.1.3). Insgesamt gelten Studien, die die Datenrezeption und -interpretation von Lehrpersonen, insbesondere auf der Mikroprozessebene adressieren, als Forschungsdesiderat (Goffin et al., 2022; Hebbecker et al., 2022; Schildkamp, 2019).

Aufgrund der Wichtigkeit von *data literacy* von Lehrpersonen, den empirischen Hinweisen auf eher geringe bis moderate Ausprägungen bei Lehrkräften und der bisher wenig vorhandenen strukturell implementierten Adressierung in Lehramtscurricula gewinnt Forschung zur Förderung von *data literacy* von angehenden Lehrpersonen an Relevanz (vgl. Kapitel 3.2.2). Allerdings liegen hier bisher nur wenige Studien vor. Dies gilt insbesondere für experimentelle Studien, die Aussagen über kausale Effekte erlauben.

Für datengestützte Entscheidungen im schulischen Kontext ergeben sich immer wieder Schnittstellen zur Digitalisierung, die in den vorherigen Kapiteln immer wieder angeklungen sind. Die mit der Digitalisierung einhergehende verstärkte und erleichterte Verfügbarkeit von Daten etwa bringt dabei aber nicht automatisch mit sich, dass diese Daten für die Gestaltung und Entwicklung von Schule und Unterricht sinnvoll genutzt werden (können), sie (positive) Wirkungen haben oder sich diese Wirkungen einfacher ergeben (Schildkamp, 2019). Da noch sehr wenige Arbeiten die Schnittstelle zwischen *data literacy* und Digitalisierung adressieren (vgl. Tabelle 1, Kapitel 3.1), stellt sich auf einer konzeptuellen Ebene die Frage, was dies für die Kompetenzen von Lehrpersonen bedeutet.

4.2. Forschungsfragen

Vor dem Hintergrund des dargestellten Forschungsstandes und der herausgearbeiteten Desiderate gilt das Forschungsinteresse der vorliegenden Arbeit der Datenrezeption und -interpretation von Lehrpersonen bei datengestützten Entscheidungen und fokussiert folgende konkrete Forschungsfragen:

1. Wie rezipieren und interpretieren Lehrpersonen Daten bei datengestützten Entscheidungen?
2. Wie können die Datenrezeption und -interpretation bei angehenden Lehrpersonen gefördert werden?
3. Welche Rolle spielt *data literacy* von Lehrpersonen im Kontext zunehmender Digitalisierung für datengestützte Entscheidungen?

4.3. Überblick über die Artikel und durchgeführten Studien

Artikel 1 (*Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien*, Kapitel 5) adressiert Forschungsfrage 1. Dabei wurde mithilfe von lautem Denken explorativ auf einer Mikroprozessebene in zwei Teilstudien untersucht, wie komplex Lehrpersonen und Lehramtsstudierende Rückmeldungen nach Vergleichsarbeiten rezipieren und interpretieren und inwiefern sich Zusammenhänge zur allgemeinen *data literacy* gemessen in einem Test zeigen. Die zwei Teilstudien wurden durchgeführt, um dabei auch die Rolle von Kontextwissen beleuchten zu können: Daher werden in Teilstudie 1 mit Lehramtsstudierenden, die kein Kontextwissen zu den VERA-Ergebnissen haben (sollten), generische, d.h. kontextunspezifische Rezeptionsprozesse untersucht, und in Teilstudie 2 kontextspezifische Rezeptionsprozesse bei Lehrpersonen aus der Praxis auf der Grundlage der VERA-Rückmeldungen der eigenen Klassen.

In Artikel 2 (*How do teachers make sense of technology-based formative assessments? Results of process mining of think-aloud data*, Kapitel 6) wird ebenso Forschungsfrage 1 auf der Basis von lautem Denken empirisch untersucht. Dabei wurde, mit einem starken Fokus auf ökologische Validität und einer Prozessperspektive, exploriert, wie Lehrpersonen Ergebnisse eigener Klassen eines technologiebasierten formativen Assessmentsystems rezipieren und interpretieren. Das spezifische Forschungsinteresse dieser Studie gilt erstens den adressierten Aspekten und Prozessschritten, die sich in den Verbalisierungen der Lehrkräfte zeigen, zweitens der Identifizierung unterschiedlicher Gruppen von Lehrkräften und drittens der Explorierung typischer Prozesse.

Artikel 3 (*Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers*, Kapitel 7) nimmt Forschungsfrage 2 in den Blick. Dabei wird untersucht, wie sich eine Onlineintervention auf die *data literacy* von Lehramtsstudierenden mit dem Fokus auf die Datenrezeption und -interpretation auswirkt. Daneben werden zudem die Effekte der Intervention auf die motivationalen Überzeugungen der Studierenden bezüglich datengestützter Entscheidungen im schulischen Kontext untersucht.

Artikel 4 (*Data-based decision making in einer digitalen Welt: Data literacy von Lehrpersonen als notwendige Voraussetzung*, Kapitel 8) adressiert Forschungsfrage 3 aus einer konzeptuellen Perspektive. Es wird für die These argumentiert, dass die *data literacy* von Lehrkräften eine notwendige Voraussetzung dafür ist, um die Potenziale für die datenbasierte Unterrichtsgestaltung und -entwicklung, die mit der Digitalisierung im Bildungswesen einhergehen, ausschöpfen und potenzielle dysfunktionale Wirkungen minimieren zu können. Dabei wird anhand aktueller digitaler Innovationen beispielhaft veranschaulicht, welche zentrale Rolle die *data literacy* für Lehrpersonen spielt, damit sich deren Potenziale bezüglich einer verstärkten Realisierung individueller Förderung bezogen auf fachliches Lernen entfalten können.

In den folgenden vier Kapiteln (Kapitel 5-8) werden nun die skizzierten Forschungsarbeiten dieser Arbeit anhand der einzelnen Artikel dargestellt. In Kapitel 9 erfolgt dann die Gesamtdiskussion der vorliegenden Arbeit.

5. Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien (Artikel 1)

Bez, S., Poindl, S., Bohl, T., & Merk, S. (2021). Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien. *Zeitschrift für Pädagogik*, *67*(4), 551–572. https://doi.org/10.3262/ZP2104551

Stichworte
Vergleichsarbeiten, Rezeption, Datenkompetenz, Think-Aloud-Methode, datenbasierte Unterrichtsentwicklung

Abstract
Im zyklischen Gesamtprozess datenbasierter Unterrichtsentwicklung gilt die Datenrezeption als notwendige Gelingensbedingung und gleichzeitig als Forschungsdesiderat. Daher untersucht die vorliegende Studie mithilfe von Think Aloud-Interviews, wie Lehrpersonen und Lehramtsstudierende Rückmeldungen von Vergleichsarbeiten rezipieren. Es zeigte sich, dass sowohl Lehrpersonen als auch Lehramtsstudierende hauptsächlich direkt in den Grafiken gegebene Entitäten rezipierten und diese vergleichend gegenüberstellten, während komplexere Elaborationen kaum vorkamen. Zudem wurde fast ausschließlich die Perspektive einer sozialen Bezugsnorm eingenommen. Eine Assoziation generischer Datenkompetenz mit der Komplexität der Elaborationen in den Think-Aloud-Protokollen zeigte sich nicht durchgehend. Zudem sprechen Bayes Faktoren gegen die Konsistenz zwischen der Datenrezeption und den Schlussfolgerungen hinsichtlich der Grafiken und Bezugsnormen.

5.1. Einleitung

Im Rahmen der sogenannten Neuen Steuerung im Bildungswesen (Altrichter & Maag Merki, 2016) wurden in allen Bundesländern Vergleichsarbeiten als ein Instrument einer Qualitätssicherung und Qualitätsentwicklung mit dem primären Ziel der datenbasierten Schul- und Unterrichtsentwicklung auf Ebene der Einzelschulen eingeführt (KMK, 2016b, 2018). Dahinter steht die Idee, dass durch die Bereitstellung und Rückmeldung von objektivierten Informationen zum Leistungsstand von Schüler*innen Prozesse auf der Ebene von Lehrpersonen und Einzelschulen angestoßen und durchgeführt werden, die zur Weiterentwicklung von Schule und Unterricht und infolge dessen zu verbesserten (fachlichen) Leistungen von Schüler*innen führen (Altrichter et al., 2016). Übersichtsarbeiten, die die bisherige Forschung seit der Einführung von Vergleichsarbeiten zu deren Rezeption und Nutzung durch Lehrpersonen zusammenfassen, sehen den forschungsmethodischen Schwerpunkt der Untersuchungen bei quantitativ ausgerichteten retrospektiven Fragebogenstudien, die auf Selbstauskünften von Lehrpersonen und Schulleitungen basieren (Altrichter et al., 2016; Dedering, 2011). Inhaltlich werden einerseits die wahrgenommene Akzeptanz, Nützlichkeit, Informativität und Verständlichkeit der Rückmeldungen sowie andererseits der selbstberichtete Umgang mit den Ergebnissen fokussiert (Altrichter et al., 2016; Dedering, 2011; Maier & Kuper, 2012).

Die zentralen Ergebnisse der Studien lassen sich folgendermaßen skizzieren: Es kann von einer grundsätzlichen Akzeptanz von und Offenheit gegenüber Vergleichsarbeiten bei Lehrkräften ausgegangen werden, wenngleich ein nicht unerheblicher Teil gewisse Skepsis hegt (Maier et al., 2012; Schliesing, 2017). Die Rückmeldungen werden insgesamt als nützlich und informativ wahrgenommen (Dedering, 2011; Schneewind, 2007) und die Verständlichkeit positiv bewertet (Koch, 2011; Maier et al., 2012; Schneewind, 2007). Geht es um konkrete Maßnahmen, die aufgrund der Testergebnisse von Lehrkräften ergriffen werden, so sind diese eher spärlich und unsystematisch und beziehen sich hinsichtlich der Unterrichtsentwicklung weniger auf grundlegende Veränderungen als vielmehr auf die Intensivierung bisheriger Praktiken und die Weiternutzung der Aufgaben im Unterricht (Altrichter et al., 2016; Groß Ophoff, 2013b; Schliesing, 2017). Als eigentlicher Ort der Analyse und Reflexion werden konzeptionell die Fachkonferenzen adressiert (Peek & Dobbelstein, 2006), allerdings belegen empirische Befunde dies nicht unbedingt (Maier et al., 2012; Wurster & Richter, 2016). Klare Evidenz für die intendierte positive Wirkung auf die (fachlichen) Leistungen der Schüler*innen liegt bislang nicht vor (Dedering, 2011; Hellrung & Hartig, 2013; Wurster et al., 2017). Insgesamt gelten Datenquellen und Forschungsdesigns jenseits retrospektiver Selbstauskünfte wie etwa die Beobachtung von Lehrpersonen als Forschungsdesiderate (Altrichter et al., 2016; Dedering, 2011), um bspw. die Datenrezeption und Ableitung pädagogischer Konsequenzen sowie deren Umsetzung bei Lehrpersonen direkt und unverzerrter zu erfassen.

Vor diesem Hintergrund fokussiert die vorliegende Studie (Mikro-)Prozesse von Lehramtsstudierenden und Lehrpersonen bei der Rezeption von Ergebnissen aus Vergleichsarbeiten: Mithilfe zweier Think-Aloud-Studien werden kognitive Elaborationen von Lehramtsstudierenden und Lehrpersonen bei der Rezeption von Rückmeldungen von Vergleichsarbeiten hochauflösend erfasst, d.h. ihre verbalisierten Gedanken und Überlegungen während der lesenden Auseinandersetzung mit den Ergebnissen. Anhand dieser Primärdaten wird dann untersucht, wie (komplex) die (angehenden) Lehrpersonen die statistischen Daten und Inhalte der Rückmeldungen rezipieren, wie konsistent die abgeleiteten Schlussfolgerungen für eigenes zukünftiges unterrichtliches Handeln aus der Datenrezeption ist und inwiefern sich Zusammenhänge zur allgemeinen Datenkompetenz zeigen.

5.2. Theoretischer Hintergrund und Forschungsstand

5.2.1. Modelle datenbasierter Unterrichtsentwicklung, der Datenkompetenz und *graph literacy*

Den Prozess der datenbasierten Unterrichtsentwicklung auf der Ebene von Lehrpersonen aus theoretischer Sicht beschreiben Helmke und Hosenfeld (2005) in ihrem Modell zur pädagogischen Nutzung von Evaluationsdaten in der Schule und unterscheiden hierbei die Schritte Rezeption (Verständnis der Daten), Reflexion (Generierung von Erklärungen) und Aktion (Umsetzung konkreter Maßnahmen), gefolgt von einer Evaluation (Überprüfung der Maßnahmenwirkungen). Im Modell werden diese Schritte durch individuelle, schulische und externe Faktoren beeinflusst. Auch internationale Rahmenmodelle und Konzeptualisierungen zu *data-based decision-making* oder *data literacy for teachers* sind zyklisch angelegt (Chick & Pierce, 2013; Mandinach & Gummer, 2016; Schildkamp, 2019), wodurch deutlich wird, dass der gelingenden Datenrezeption, d.h. dem adäquaten Verstehen der statistischen Informationen, entscheidende Bedeutung zukommt: Sie steht am Beginn des Prozesses und stellt daher eine notwendige Voraussetzung für eine treffende Interpretation der Ergebnisse unter Berücksichtigung von Kontextinformationen und für die sich anschließende angemessene Ableitung und Umsetzung geeigneter (Unterrichts-)Maßnahmen dar. Auch wenn die Modelle (Chick & Pierce, 2013; Coburn & Turner, 2011; Marsh, 2012; Schildkamp, 2019) zwar unterschiedliche Begrifflichkeiten und Foki verwenden, ist ihnen gemeinsam, dass zwischen der eigentlichen Datenanalyse bzw. -rezeption (*noticing*, *reading*, *analyzing*), und der Interpretation, also der Reflexion der Daten unter Rückbindung an den spezifischen Kontext (*interpreting, sense making*, *combining expertise to build knowledge*), unterschieden wird. Datenkompetenz bei Lehrkräften bzw. *data literacy for teachers* bezieht sich damit nicht nur auf das alleinige Verstehen von (statistischen) Daten sondern kann nach Mandinach und Gummer erfasst werden als

> “the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data [...] to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn.” (Mandinach & Gummer, 2016, S. 367)

Rückmeldungen nach Vergleichsarbeiten weisen neben allgemeinen Informationen und Hinweisen hauptsächlich grafisch aufbereitete Testergebnisse mit unterschiedlichen Abstraktions- und Aggregatebenen auf. Die Rezeption der Rückmeldungen, also die konkreten kognitiven Elaborationen im Zuge der lesenden Auseinandersetzung, lässt sich daher am besten als eine Subkomponente einer allgemeinen Datenkompetenz mit dem Konzept der *graph comprehension* bzw. *graph literacy* erfassen. Diese lässt sich definieren

als die Fähigkeit, grafisch repräsentierte Informationen zu lesen und zu verstehen (Friel et al., 2001; Galesic & Garcia-Retamero, 2011). In der Literatur werden typischerweise drei unterschiedliche Niveaustufen konzeptualisiert (Friel et al., 2001; Galesic & Garcia-Retamero, 2011; Koch, 2013; van den Bosch et al., 2017; van den Hurk et al., 2016; Zeuch et al., 2017), die auch der vorliegenden Studie zugrunde gelegt werden: *Reading the data* (unterste Stufe) beschreibt das Extrahieren explizit encodierter Entitäten, d.h. das Ablesen einzelner direkt gegebener Datenpunkte. *Reading between the data* (mittlere Stufe) umfasst das Herstellen von Beziehungen zwischen Daten oder Datenpunkte zu neuen Kategorien zusammenzufassen. *Reading beyond the data* (höchste Stufe) meint das Zusammenfassen der Grafik in einer Gesamtaussage, die Generierung neuer statistischer Entitäten, Schlüsse zu ziehen oder Vorhersagen aufgrund der Daten in der Grafik zu treffen. Dabei stellen der Grad der Aggregation (einzelne Datenpunkte vs. alle Datenpunkte) sowie das Ausmaß der Rückbindung der Daten an einen spezifischen Kontext (gar nicht vs. relativ hoch) zentrale Kriterien zur Abgrenzung der Stufen dar. Im Modell zur Datenkompetenz von Chick und Pierce (2013) werden diese Niveaustufen (dort benannt als *technical skills*) zudem flankiert durch das Kontextwissen einerseits zu lokalen Bedingungen (z.B. dem Hintergrund der Schüler*innen) und andererseits zu dem Instrument, mit dem die Daten generiert werden (z.B. grundlegendes Wissen zu Vergleichsarbeiten).

5.2.2. Datenkompetenz und *graph literacy* bei Lehrpersonen und Lehramtsstudierenden

Der kompetente Umgang mit statistischen Daten ist für die datenbasierte Unterrichtsentwicklung äußerst relevant, wobei bezweifelt werden kann, inwiefern dieser bei Lehrkräften hinreichend ausgeprägt ist bzw. vorausgesetzt werden kann (Altrichter et al., 2016; Maier et al., 2012; Peek & Dobbelstein, 2006; Zimmer-Müller et al., 2014): Lehrpersonen gelten, mitbedingt durch das Lehramtsstudium, das in der Regel keine oder nur begrenzte forschungsmethodische bzw. statistische Anteile beinhaltet (Stelter & Miethe, 2019), als statistische Laien. Lehramtsstudierende aus dem MINT-Bereich zeigen insgesamt eine höhere allgemeine Datenkompetenz als Studierende aus anderen Fächerbereichen (Merk et al., 2020) und Mathematiklehrkräfte schneiden auch in *graph literacy*-Tests signifikant besser ab (Zeuch et al., 2017), was durch entsprechende Fachstudienanteile erklärt werden kann. In den Rezeptionsstudien zu Vergleichsarbeiten wurde zwar die eingeschätzte Verständlichkeit der Rückmeldungen durch Lehrpersonen mit erfasst. Allerdings muss diese vom tatsächlichen Verstehen der Daten in den Rückmeldungen durch Lehrpersonen unterschieden werden (Schliesing, 2017). Diejenigen (wenigen) Studien, die die tatsächliche Datenkompetenz bei Lehrkräften untersuchen, zeigen, dass Lehrkräfte nicht unbedingt über eine hohe Datenkompetenz verfügen und Schwierigkeiten beim Verstehen und im Umgang mit Grafiken,

auch mit grafischen Darstellungen aus Vergleichsarbeiten, haben, obwohl sie sie als verständlich einschätzen (Koch, 2011; Schliesing, 2017). Zum Einfluss von Kontextwissen (Chick & Pierce, 2013) gibt es zwar erste Studien, die die Rolle von Kontextwissen auf *beliefs*, den selbsteingeschätzten Umgang mit den Daten, die Selbstwirksamkeit sowie die Angst bezüglich der Daten adressieren (z.B. Reeves & Chiang, 2018). Allerdings besteht hier weiterer Forschungsbedarf, inwiefern sich Kontextwissen auf Rezeptionsprozesse auswirken kann.

5.2.3. Relevanz der Bezugsnormen

Je nach zuständigem Institut unterscheiden sich die Formate von Vergleichsarbeiten zwar voneinander (Groß Ophoff, 2013b; Zimmer-Müller et al., 2014). Generell aber bestehen sie aus Ergebnisdarstellungen, in denen neben den Kompetenzstufenzuordnungen auf Klassen- und Individualebene sowie Lösungshäufigkeiten der einzelnen Aufgaben und Leitideen auch Ergebnisse der Einzelschule und Landesergebnisse der jeweiligen Schulart bzw. einer Kontextgruppe (fairer Vergleich) für die jeweiligen im Schwerpunkt geprüften Kompetenzbereiche dargestellt sind (Tarkian et al., 2019). Damit können Lehrkräfte die Ergebnisse ihrer Klasse sowohl sozial als auch kriterial normieren. Auch eine ipsative (d.h. individuelle) Bezugsnorm (Rheinberg, 2011) ist möglich, wenn Ergebnisse mit Ergebnissen aus anderen Kompetenzbereichen oder Leitideen (zum selben Zeitpunkt erfasst) verglichen werden. Bezugsnormen sind theoretisch bedeutsam für die Ableitung von Handlungsmaßnahmen: So impliziert eine ipsative Perspektive etwa curriculare Verschiebungen zwischen Leitideen und eine kriteriale (z.B. an den Kompetenzstufen orientierte) Perspektive eher Veränderungen für das Unterrichten innerhalb eines Kompetenzbereichs.

5.2.4. Fragestellung und Forschungsfragen

Die vorliegende Studie fokussiert anlehnend und ergänzend zu den bisherigen Forschungsarbeiten explorativ in zwei Teilstudien die kognitiven Elaborationen von Lehrkräften bei der Rezeption von VERA-Ergebnissen mithilfe der Think-Aloud-Methode (Ericsson & Simon, 1998; Espin et al., 2017; Leighton, 2017; Padilla & Leighton, 2017; van Someren et al., 1994) mit der übergeordneten Fragestellung: Wie werden Rückmeldungen aus Vergleichsarbeiten rezipiert und inwiefern zeigen sich Zusammenhänge zwischen der Datenkompetenz der Personen und ihren Rezeptionsprozessen in der Performanz der Think-Aloud-Protokolle? Diese übergreifende Fragestellung wurde differenziert in zwei Teilstudien untersucht, wobei die oben beschriebene Rolle von Kontextwissen berücksichtigt wurde: In Teilstudie 1 wurden generische (kontextunspezifische) Rezeptionsprozesse von

Lehramtsstudierenden bzgl. VERA-Rückmeldungen und in Teilstudie 2 kontextspezifische Rezeptionsprozesse von Lehrpersonen bzgl. eigener VERA-Rückmeldungen untersucht.

Studie 1

Lehramtsstudierende haben in der Regel keine Erfahrung mit Vergleichsarbeiten bzw. dem Umgang mit Rückmeldungen und verfügen so über kein Erfahrungs- und Kontextwissen, das für die Rezeption bedeutsam sein könnte. Ebenso scheint fraglich, ob sie bereits über das notwendige pädagogische, fachliche und fachdidaktische Wissen verfügen, um ausgehend von Ergebnisdarstellungen nach Vergleichsarbeiten unterrichtliche Handlungsmaßnahmen konstruieren zu können (Reeves & Chiang, 2018). Daher fokussiert diese Teilstudie auf generische Rezeptionsprozesse und untersucht die beschriebene Fragestellung mit den folgenden Forschungsfragen:

1. Welche Informationen und statistischen Entitäten, die in den VERA-Rückmeldungen enthalten sind, rezipieren Lehramtsstudierende?
2. Inwiefern zeigen datenkompetente Lehramtsstudierende komplexere und korrektere Elaborationen bei der Rezeption der Rückmeldungen?

Studie 2

Lehrpersonen aus der schulischen Praxis hingegen sind mit der Durchführung von Vergleichsarbeiten und daher auch mit den Rückmeldungen vertraut. Sie verfügen zudem über vielfältiges Kontextwissen hinsichtlich des eigenen Unterrichts, ihrer Schüler*innen, spezifischer Bedingungen ihrer Einzelschule etc., was für die Rezeption, Interpretation und die mögliche Ableitung von Handlungsmaßnahmen bedeutsam scheint (Chick & Pierce, 2013). Bei dieser Teilstudie mit Lehrpersonen sind daher die folgenden Forschungsfragen leitend:

1. Wie komplex rezipieren Lehrpersonen die VERA-Rückmeldungen ihrer eigenen Klassen und welche Bezugsnormen adressieren sie dabei?
2. Inwiefern zeigt sich Konsistenz zwischen der Rezeption der Ergebnisse und den abgeleiteten Schlussfolgerungen für unterrichtliches Handeln hinsichtlich der adressierten Grafiken einerseits und der Bezugsnormen andererseits?
3. Inwiefern zeigen datenkompetente Lehrpersonen komplexere Elaborationen bei der Rezeption ihrer Rückmeldungen?

5.3. Methode[1]

5.3.1. Stichprobe

An Teilstudie 1 nahmen *N* = 76 Lehramtsstudierende (*w* = 58%) für den Sekundarbereich im Rahmen ihres Bildungswissenschaftlichen Studiums teil. Alle Studierenden waren am selben Hochschulort immatrikuliert. 52% waren zwischen 21 und 23 Jahre alt (24-27 Jahre: 38%). 55% befanden zwischen dem 7. und 9. Semester (4.-6. Sem. 23%, 10.-12. Sem. 20%).

Teilstudie 2 basiert auf einer Stichprobe von *N* = 25 (*w* = 68%) Lehrpersonen aus dem Primar- und Sekundarbereich in Baden-Württemberg und schließt sowohl Lehrkräfte aus unterschiedlichen Fächern ein als auch Lehrkräfte, die VERA 3 und VERA 8 durchgeführt hatten. 32% der Lehrpersonen waren zwischen 32 und 37 Jahre alt (27-31 Jahre: 28%; 38-43 Jahre: 24%). 40% verfügten über 1 bis 4 Jahre Berufserfahrung (24%: >14 Jahre; 20%: 5-9 Jahre).

5.3.2. Design und Ablauf

Studie 1

Um zu untersuchen, wie Lehramtsstudierende Rückmeldungen rezipieren und welche Entitäten und (Teil-)Informationen sie adressieren, wurden die Studierenden mithilfe der Think-Aloud-Methode gebeten, während der Rezeption zweier VERA-Grafiken einen Screencast ihrer Äußerungen aufzunehmen und dabei auf die adressierten Stellen zu zeigen. Alle Studierenden erhielten dazu dieselben zwei Grafiken einer authentischen VERA-Rückmeldung aus Baden-Württemberg; zum einen eine Grafik zur Kompetenzstufenverteilung und zum anderen eine Grafik zur Darstellung von aufgabenspezifischen Lösungshäufigkeiten (siehe Abbildung 5 und 6). Anschließend bearbeiteten die Studierenden einen Datenkompetenztest (siehe Abschnitt 5.3.3). Die audiovisuellen Daten der Think-Aloud-Protokolle wurden mithilfe deduktiv entwickelter Kategorien hinsichtlich der Nennung und der Korrektheit bestimmter Grafikinhalte und Rezeptionsschritte ausgewertet (siehe Abschnitt 5.3.4).

[1] Materialien, Datensätze sowie die reproduzierbare Dokumentation der Datenanalyse sind auf dem Open Science Framework unter https://osf.io/xmafk/ verfügbar.

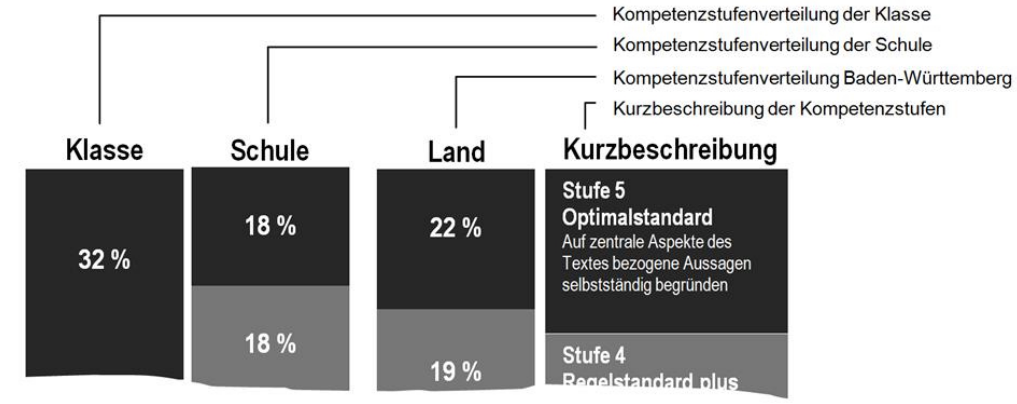


Abbildung 5. Grafik 1: Kompetenzstufengrafik (Ausschnitt) (© Institut für Bildungsanalysen Baden-Württemberg (IBBW), 2019, S. 7, Abdruck mit freundlicher Genehmigung)

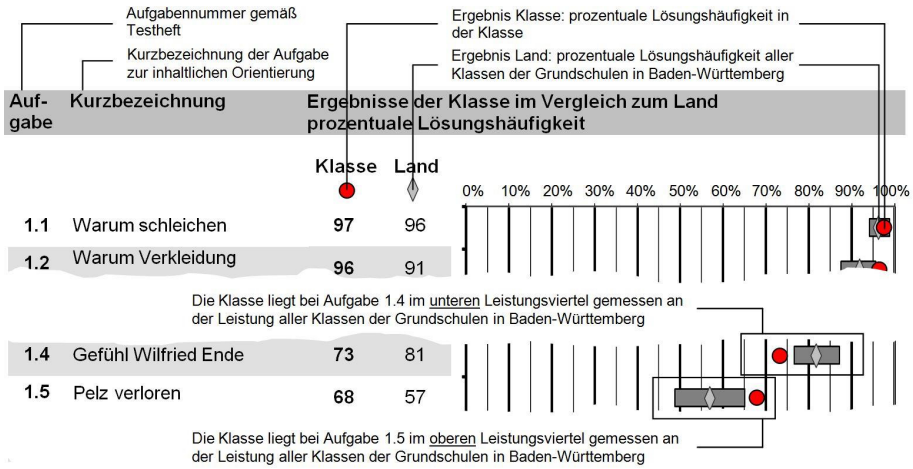


Abbildung 6. Grafik 2: Lösungshäufigkeiten der einzelnen Aufgaben (Ausschnitt) (Institut für Bildungsanalysen Baden-Württemberg (IBBW), 2019, S. 8, Abdruck mit freundlicher Genehmigung)

Studie 2

Um Daten zu den Rezeptionsprozessen von Lehrkräften, die in ihrem Berufsalltag mit Vergleichsarbeiten und deren Rückmeldungen konfrontiert sind, zu generieren, wurden die Lehrpersonen zunächst gebeten, ihre Gedanken und Überlegungen während der Betrachtung der Ergebnisrückmeldung ihrer Klasse laut zu äußern und auf die entsprechenden Stellen zu zeigen. Dann wurden folgende Impulse gegeben: „Gibt es Schlüsse, die Sie aus den VERA-Ergebnissen ziehen?“ und „Inwiefern sind die dargestellten VERA-Ergebnisse für Ihr unterrichtliches Handeln nützlich oder informativ?“. Die Lehrpersonen nutzten dazu aktuelle Rückmeldungen ihrer Klasse. Sechs Lehrpersonen bearbeiteten keine eigene sondern eine andere Rückmeldung, da sie im Zeitraum der Studie keine Vergleichsarbeiten mit eigenen Klassen durchgeführt hatten. Im zweiten Teil bearbeiteten die Lehrpersonen einen

Datenkompetenztest. Die audiovisuellen Daten der Think-Aloud-Protokolle wurden von zwei geschulten Rater*innen unabhängig voneinander mithilfe eines deduktiv-induktiv entwickelten Schemas hinsichtlich der Komplexität der Datenrezeption (*graph literacy*), der Bezugsnormen sowie der Schlussfolgerungen für das eigene unterrichtliche Handeln geratet (siehe Abschnitt 5.3.4).

5.3.3. Instrumente

Zur Erfassung der Datenkompetenz der Lehramtsstudierenden und Lehrpersonen bearbeiteten sie adaptierte Kurztests, deren Langversion von Merk et al. (2020) anlehnend an die theoretische Konzeptualisierung von Datenkompetenz von Lehrkräften (*data literacy for teachers,* DLFT) von Mandinach und Gummer (2016) und unter Einbezug von Items aus einem validierten Instrument von Koch (2013) entwickelt und validiert wurde. Er bildet die Inhaltsbereiche *use data* und *transform data into information* der DLFT ab (Mandinach & Gummer, 2016). Die Eindimensionalitätsannahme wurde durch die Ergebnisse einer konfirmatorischen Faktorenanalyse basierend auf Diagonally-Weighted- Least-Square-Schätzern (DWLS), robusten Standardfehlern und Teststatistiken gestützt ($\chi^2(35) = 42.8$, Confirmatory-Fit-Index [CFI] = .985, Tucker-Lewis-Index [TLI] = .978, Root-Mean-Square-Error-of-Approximation [RMSEA] = .073). Die interne Konsistenz wurde basierend auf tetrachorischen Korrelationen geschätzt (ordinales Cronbach's α; Gadermann et al., 2012) und zeigte gute Ergebnisse (Studie 1: α = .767; Studie 2: α = .737).

5.3.4. Auswertung der Think-Aloud-Protokolle

Studie 1

Lehramtsstudierende haben in der Regel keine Erfahrung im Umgang mit Ergebnisrückmeldungen. Daher wurde zur Auswertung der Think-Aloud-Protokolle ein grafikspezifisches deduktives Kategorienschema von niedriger bis mittlerer Inferenz entwickelt, um zu erfassen, welche Entitäten und (Teil-)Informationen wie etwa das Ablesen der Prozentwerte bei einzelnen Kompetenzstufen oder Lösungshäufigkeiten der Klasse bei einzelnen Leitideen im Vergleich zu den Landesergebnissen die Lehramtsstudierenden in den Grafiken adressieren (Schema verfügbar unter https://osf.io/xmafk/). Die Think-Aloud-Äußerungen der Lehramtsstudierenden wurden von zwei geschulten Rater*innen mithilfe dieses Schemas hinsichtlich der Nennung und Korrektheit unabhängig voneinander beurteilt. Die Interraterreliabilitätsprüfung mithilfe von Krippendorffs α (Hayes & Krippendorff, 2007) ergab mit einer Ausnahme befriedigende bis sehr gute Werte ($.55 \leq \alpha \leq .94$; Ausnahme: $\alpha \leq .39$.) Alle Nicht-Übereinstimmungen wurden in Konsensurteile überführt.

Studie 2

Lehrpersonen aus der Praxis sind mit der Durchführung von VERA betraut, sollen diese für ihre Unterrichtsentwicklung nutzen und verfügen so über vielfältiges Kontextwissen. Daher wurde für die Rezeption der eigenen VERA-Rückmeldungen bei Lehrkräften ein höher inferentes Ratingschema entwickelt, das neben der Komplexität der Datenrezeption und den Bezugsnormen auch eine Erfassung der Schlussfolgerungen für das eigene unterrichtliche Handeln (Forschungsfrage 2) ermöglicht. Dazu wurden zunächst deduktiv für das Rating der Komplexität der Datenrezeption die drei Niveaustufen der *graph literacy, reading the data, reading between the data* und *reading beyond the data* (Friel et al., 2001; Galesic & Garcia-Retamero, 2011) zugrunde gelegt, zwischen sozialer, kriterialer und ipsativer Bezugsnorm unterschieden sowie eine allgemeine Kategorie für Schlussfolgerungen für weitere Unterrichtsprozesse gesetzt. Da die Lehrpersonen vorrangig nur zwei Darstellungsarten der unterschiedlichen Grafiken der Rückmeldungen rezipierten, wurden nur diese weiter ausgewertet. Es handelte sich dabei (wie in Studie 1) um die Darstellungen der Kompetenzstufenverteilungen (Grafik 1) und der Lösungshäufigkeiten zu einzelnen Aufgaben (Grafik 2) zu den jeweils geprüften Kompetenzbereichen. Im Zuge einer zweiten induktiven Überarbeitung des Auswertungsschemas wurde es inhaltlich geschärft, *reading between the data* und *reading beyond the data* wegen nicht zufriedenstellender Interraterreliabilitätswerte wiederholt geratet und die Kategorie *Schlussfolgerungen* in die abstrakte Formulierung eines allgemeinen Handlungsbedarfs einerseits und konkreter Implikationen für das künftige unterrichtliche Handeln andererseits konkretisiert (Schema verfügbar unter https://osf.io/xmafk/). Die audiovisuellen Daten wurden anhand dieses Schemas als *timed-event codings* (Bakeman & Quera, 2011) von geschulten Rater*innen unter Prüfung der Interraterreliabilität nach Krippendorffs α (Hayes & Krippendorff, 2007) unabhängig voneinander geratet und alle Nicht-Übereinstimmungen in Konsensurteile überführt. Die Interraterreliabilität ergab befriedigende bis sehr gute Werte ($.58 \leq \alpha \leq .90$).

5.4. Ergebnisse

5.4.1. Studie 1: Lehramtsstudierende

Forschungsfrage 1

Auf Grundlage der Think-Aloud-Kodierungen in Studie 1 wurde bezüglich der ersten Forschungsfrage untersucht, welche Informationen Lehramtsstudierende den Grafiken entnehmen und welche komplexeren Elaborationen (wie etwa die Gruppierung von Daten) vorgenommen werden. Dabei wurde auch erfasst, ob die Kategorie in den Think-Aloud-Protokollen nur genannt wurde oder ob die Äußerung auch korrekt war. Die Ergebnisse sind

für die beiden zu rezipierenden Grafiken visualisiert dargestellt (Abbildung 7: Rezeption Grafik 1 (Kompetenzstufenverteilung); Abbildung 8: Rezeption Grafik 2 (Lösungshäufigkeiten)).

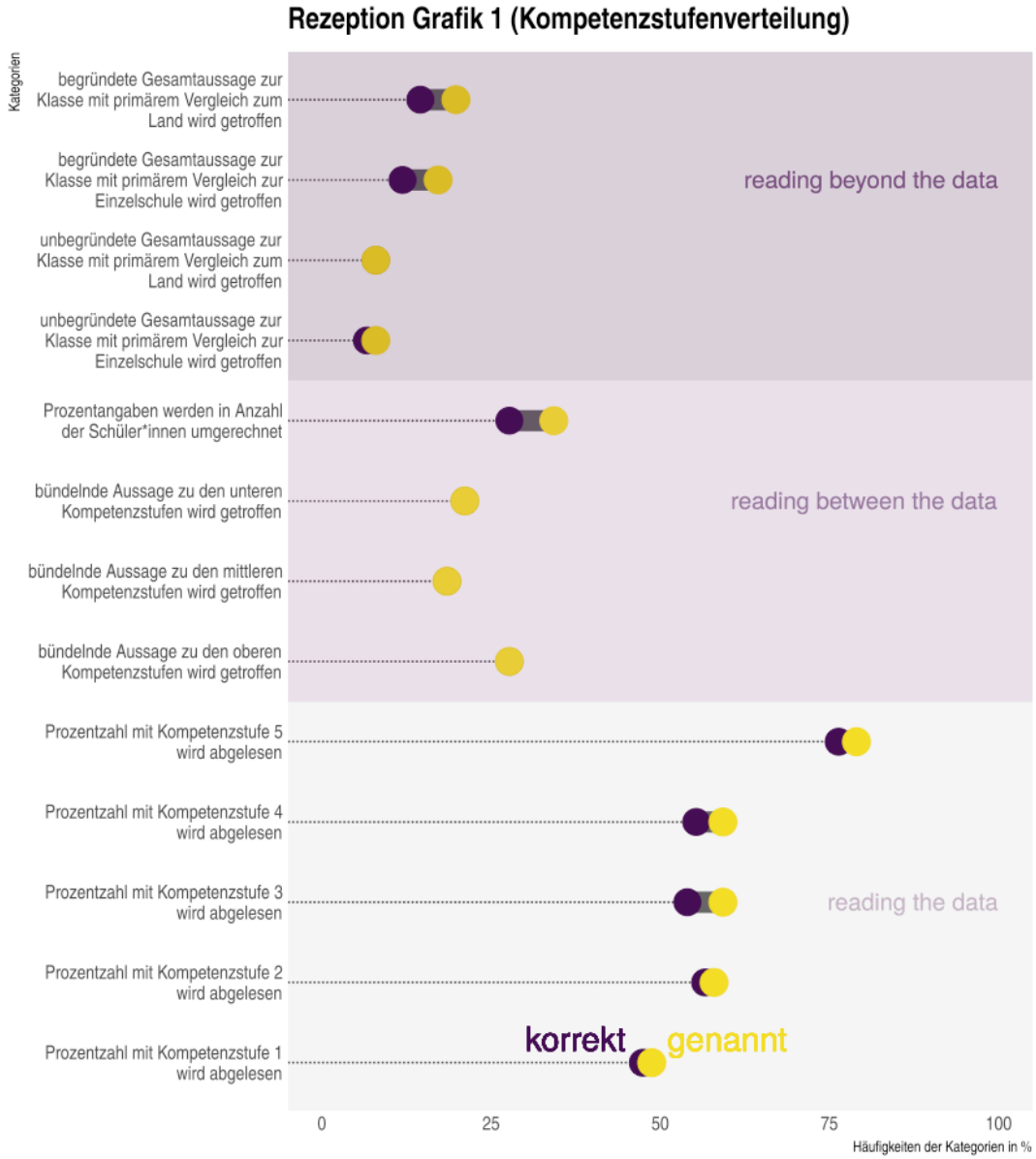


Abbildung 7. Rezeption Grafik 1 (siehe Abbildung 5)

Anmerkung. Helle Punkte markieren den Anteil der reinen Nennung, dunkle Punkte den Anteil der korrekten Äußerung einer entsprechenden Kategorie. Sind nur helle Punkte sichtbar, fallen die Häufigkeiten von Korrektheit und Nennung zusammen. Der Balken zwischen hellen und dunklen Punkten visualisiert die Differenz zwischen beiden Anteilen. Im Hintergrund sind die Zuordnungen der Kategorien, die inhaltsbezogen auf die einzelnen Grafiken sind, zu den konzeptionellen Stufen der graph literacy (reading the data, reading between the data, reading beyond the data) eingefärbt.
Bsp. Kategorie „Prozentzahl mit Kompetenzstufe 1 wird abgelesen“: Rund 50% der Studierenden lasen die Prozentzahl mit Kompetenzstufe 1 während der Think-Aloud-Protokolle ab, beinahe alle dieser Äußerungen waren auch korrekt.
Die farbige Grafik in hochauflösender Darstellung und die Wertetabelle sind verfügbar unter https://osf.io/xmafk/.

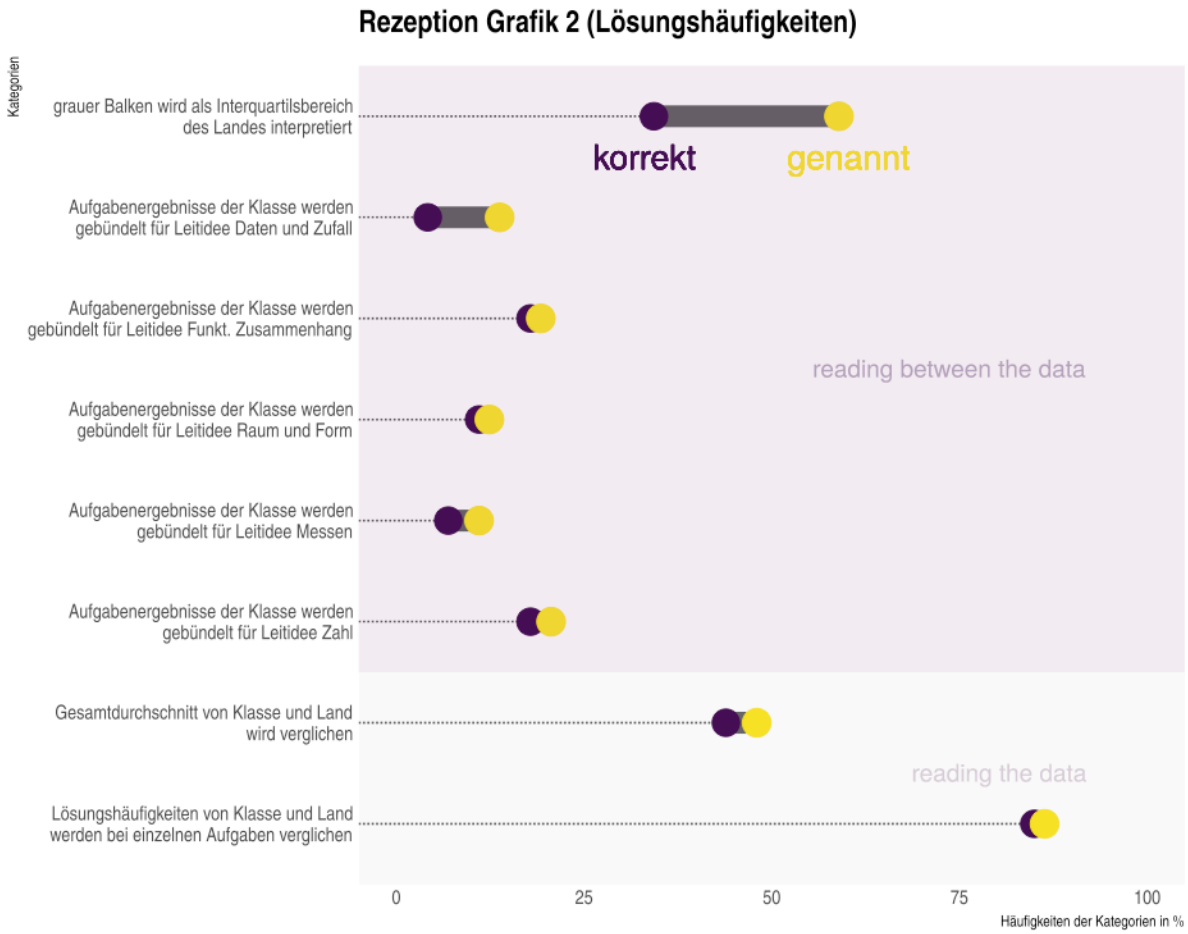


Abbildung 8. Rezeption Grafik 2 (siehe Abbildung 6)

Anmerkung. Helle Punkte markieren den Anteil der reinen Nennung, dunkle Punkte den Anteil der korrekten Äußerung einer entsprechenden Kategorie. Der Balken zwischen hellen und dunklen Punkten visualisiert die Differenz zwischen beiden Anteilen. Im Hintergrund sind die Zuordnungen der Kategorien, die inhaltsbezogen auf die einzelnen Grafiken sind, zu den konzeptionellen Stufen der graph literacy (reading the data, reading between the data, reading beyond the data) eingefärbt.
Bsp. Kategorie „grauer Balken wird als Interquartilsbereich des Landes interpretiert“: Etwas mehr als 50% der Studierenden adressierte den Interquartilsbereich des Landes, aber nur rund 30% der Studierenden interpretierte ihn korrekt.
Die farbige Grafik in hochauflösender Darstellung und die Wertetabelle sind verfügbar unter https://osf.io/xmafk/.

Abbildung 7 zeigt, dass das einzelne Ablesen direkt gegebener Informationen wie z.B. die Prozentzahl je Kompetenzstufe in Grafik 1 oder der Vergleich einzelner Lösungshäufigkeiten in Grafik 2 mit Werten von rund 50% insgesamt häufiger – und auch häufiger korrekt – auftritt als das Gruppieren von Daten in Leistungsgruppen oder das Treffen einer Gesamtaussage über die Grafik. Für Lehramtsstudierende scheint es also schwieriger zu sein, Daten zu Gruppen zu bündeln, was der mittleren Stufe *reading between the data* entspricht, oder eine Gesamtaussage zu treffen, was der höchsten Stufe *reading beyond the data* entspricht, als einzelne, direkt in der Grafik angegebene Werte abzulesen, was der niedrigsten Stufe *reading the data* zuzuordnen ist. Für Grafik 2 zeigt sich insbesondere, dass zwar über 50% der Studierenden den Interquartilsbereich der Landesergebnisse adressierte, aber die Korrektheit

weit darunter liegt. Auch dies scheint aus theoretischer Sicht plausibel, da ein konzeptionelles Verständnis des Interquartilsbereichs bereits als *reading between the data* einzuordnen ist.

Forschungsfrage 2

Um zu untersuchen, inwiefern die mithilfe des Tests gemessene Datenkompetenz komplexere Elaborationen in den Think-Aloud-Kodierungen prädiziert (Forschungsfrage 2), wurden aufgrund der deskriptiven Ergebnisse zur Häufigkeit des korrekten Auftretens und theoretischer Überlegungen diejenigen Kategorien ausgewählt, die den höheren Stufen der *graph literacy* zugeordnet werden können. Diese wurden durch die Scores der Datenkompetenztests jeweils mithilfe logistischer Regressionen prädiziert. Dabei zeigten sich für das Erreichen von *reading between the data* bei Grafik 1 ($\beta_1 = .586$, $p = .043$) sowie für die korrekte Interpretation des Interquartilsbereich des Landes ($\beta_1 = .595$, $p = .045$) positive signifikante Zusammenhänge substantieller Größe, während die Zusammenhänge für *reading beyond the data* bei Grafik 1 und *reading between the data* bei Grafik 2 nicht signifikant waren. Da nicht-signifikante *p*-Werte nicht zwischen *Absence of Evidence* und *Evidence of Absence* unterscheiden (Dienes, 2016), wurden für diese Kategorien *Adjusted Approximative Fractional Bayes Factors* (Gu et al., 2018) berechnet. Bei *reading between the data* ergab sich ein Bayes Faktor von $BF_{1c} = 10.34$. Dies bedeutet, dass die Daten 10.34-mal wahrscheinlicher unter der Annahme eines positiven Zusammenhangs als unter der Annahme des Gegenteils ($\beta_1 \leq 0$), was meinst als substanzielle Evidenz interpretiert wird. Bei *reading beyond the data* bei Grafik 1 zeigten sich inkonklusive Ergebnisse ($BF_{1c} = .615$).

5.4.2. Studie 2: Lehrpersonen

Forschungsfrage 1

Um zu untersuchen, wie komplex Lehrpersonen VERA-Rückmeldungen ihrer Klassen rezipieren und welche Bezugsnormen sie dabei adressieren (Forschungsfrage 1), wurden die Anteile der *graph literacy*-Stufen und der Bezugsnormen in den Think-Aloud-Protokollen pro Grafik und grafikübergreifend berechnet, indem die Dauer der entsprechenden Äußerungen ins Verhältnis zur Gesamtdauer pro Protokoll gesetzt wurde. Die Länge der Think-Aloud-Protokolle der Lehrkräfte betrug im Mittel 5.8 Minuten (*Min* = 2.4, *Max* = 13.8, *SD* = 2.2). Die deskriptiven Werte bezogen auf Forschungsfrage 1 sind in Tabelle 2 dargestellt. Diese Werte zeigen, dass Lehrpersonen in der Tendenz eine mittlere Komplexität in der Datenrezeption (*reading between the data*) aufweisen und hauptsächlich die soziale Bezugsnorm adressieren. Die kriteriale und vor allem die ipsative Perspektive wird kaum eingenommen.

Anteile der graph literacy-Stufen und der Bezugsnormen bei der Rezeption			
	grafik-übergreifend	**Grafik 1**	**Grafik 2**
reading the data	*Md* = 1.9 *IQR* [.0-4.8]	*Md* = .0 *IQR* [.00-2.3]	*Md* = .0 *IQR* = [.0-.0]
reading between the data	*Md* = 30.6 IQR [16.4-36.6]	*Md* = 17.8 *IQR* [12.3-27.7]	*Md* = 9.6 *IQR* = [3.1-15.8]
reading beyond the data	*Md* = 6.6 *IQR* [1.1-9.5]	*Md* = 2.1 *IQR* [.0-3.4]	*Md* = 3.8 *IQR* = [.0-6.8]
Anteil kriterialer Bezugsnorm	*Md* = 5.9 *IQR* [.0-15.9]	*Md* = 4.9 *IQR* [.0-12.2]	*Md* = .0 *IQR* = [.0-2.2]
Anteil sozialer Bezugsnorm	*Md* = 23.6 *IQR* [12.7-30.8]	*Md* = 11.8 *IQR* [6.2-20.1]	*Md* = 10.7 IQR [.0-20.2]
Anteil ipsativer Bezugsnorm	*Md* = .0 *IQR* [.0-3.4]	*Md* = .0 *IQR* [.0-.0]	*Md* = .0 *IQR* [.0-1.1]

Tabelle 2. Anteile der *graph literacy*-Stufen und der Bezugsnormen bei der Rezeption

Anmerkung. Angegeben sind jeweils Median sowie 1. und 3. Quartil der prozentualen Anteile bezogen auf die Gesamtdauer des Think-Aloud-Protokolls.
Beispiel reading between the data: Die Lehrkräfte äußerten sich im Mittel (hier: Md) während 30.6% der Interviewdauer auf der graph literacy-Stufe reading between the data. Bezieht man sich nur auf Grafik 1, beträgt dieser Anteil im Median 17.8%.

Um nicht nur die jeweiligen zeitlichen Anteile der interessierenden Kategorien in ihrer zentralen Tendenz und Verteilung abzubilden sondern auch die Mikroprozesse der einzelnen Lehrpersonen in ihrem Verlauf zu explorieren, wurden die Elaborationen je Lehrkraft anhand der Ratings in ihrem zeitlichen Verlauf visualisiert (Abbildung 9).

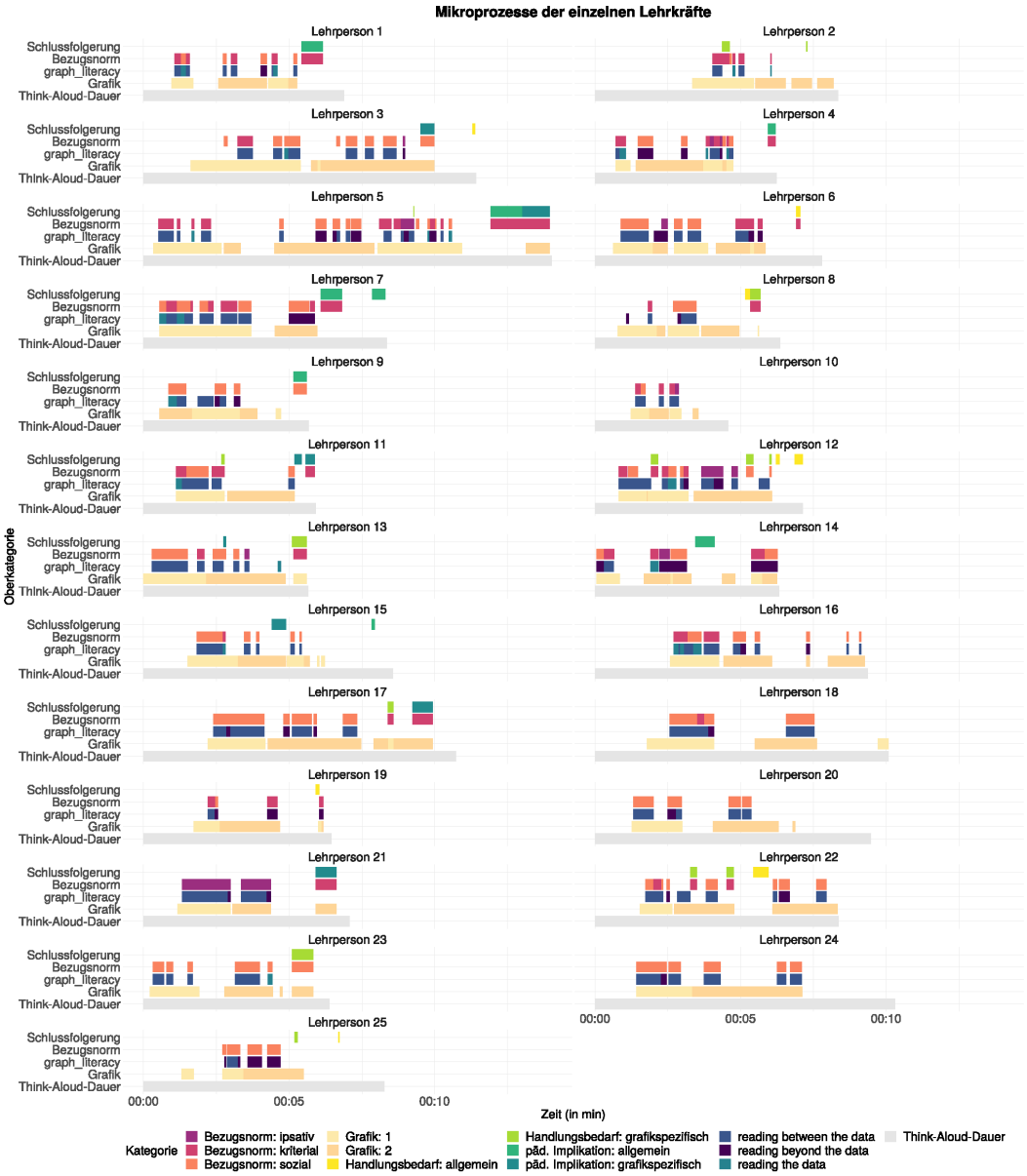


Abbildung 9. Rezeptionsverläufe der einzelnen Lehrpersonen

Anmerkung. Pro Lehrperson ist im zeitlichen Verlauf dargestellt, wie lange das Think-Aloud-Protokoll ist (Think-Aloud-Dauer), inwiefern sich die Äußerungen auf eine bestimmte Grafik beziehen (Grafik), auf welcher graph literacy-Stufe die Äußerung einzuordnen ist (graph_literacy), welche Bezugsnorm adressiert wird (Bezugsnorm) und welche Art von Schlussfolgerung die Lehrperson formuliert (Schlussfolgerung).
Bsp. Lehrperson 24: Sie rezipiert zuerst Daten auf dem Niveau reading between data und mit sozialer Bezugsnorm aus Grafik 1, trifft bei ca. 2.5min eine kurze Aussage auf dem Niveau reading beyond the data und wechselt dann wieder zu reading between the data. Ab ca. 4.30min wechselt sie zu Grafik 2, wobei hier nur die Grafik adressiert wird, aber keine Datenrezeption erfolgt, abgesehen von zwei kurzen Abschnitten mit sozialer Bezugsnorm und Datenrezeption auf dem Niveau reading between the data. Schlussfolgerungen werden nicht formuliert (anders z.B. Lehrperson 23, die gegen Ende Handlungsbedarf mit sozialer Bezugsnorm bezogen auf Grafik 2 äußert).
Die farbige Grafik in hochauflösender Darstellung ist verfügbar unter https://osf.io/xmafk/.

Anhand dieser Visualisierung wird ersichtlich, dass einzelne Lehrpersonen während ihrer individuellen Informationsgenerierung immer wieder zwischen den einzelnen Grafiken und den Bezugsnormen wechseln, während es auch Lehrpersonen gibt, die recht stabil eine einzige, häufig die soziale, Bezugsnorm adressieren. Hinsichtlich der Bezugsnormen fällt

Lehrperson 21 besonders ins Auge, da sie fast ausschließlich ipsativ während der Datenrezeption normiert. Insgesamt zeigen sich deutliche individuelle Unterschiede in den Mikroprozessen bzw. drängen sich keine eindeutigen Annahmen über Rezeptionsmuster auf.

Forschungsfrage 2

Schon bei der Visualisierung der Rezeptionsverläufe der Lehrpersonen (Abbildung 9) zeigte sich, dass bei allen Lehrkräften der Anteil der Schlussfolgerungen im Vergleich zur Datenrezeption weitaus geringer ist. In Bezug zur zweiten Forschungsfrage (Inwiefern zeigt sich Konsistenz zwischen der Datenrezeption und dem Ableiten handlungsleitender Schlussfolgerungen bezüglich der Grafiken und der Bezugsnormen?) muss festgehalten werden, dass 20 von insgesamt 25 Lehrpersonen Schlussfolgerungen für ihr zukünftiges unterrichtliches Handeln auf Basis ihrer VERA-Rückmeldungen (spontan oder auf Nachfrage) konstruierten. Dabei äußerten 13 Lehrpersonen allgemeinen Handlungsbedarf (z.B. „Generell finde ich es schonmal gut, dass so viele im Regelstandard sind, man könnte natürlich noch schauen, dass man die 20% aus dem unteren Mindeststandard noch weiter reduzieren könnte.") und 12 Personen konkrete Implikationen für ihr unterrichtliches Handeln (z.B. „Vor allem diejenigen Kinder, die bei ‚Größen und Messen' im unteren Drittel sind, da heißt es für mich, die Repräsentanten zu verinnerlichen: Das mit Material zu legen, zu messen, zu wiegen, dass sie immer mehr eine Vorstellung bekommen, dass eine Runde um den Sportplatz einfach nicht 400 Kilometer lang sein kann."). Hinsichtlich der Frage nach der Konsistenz zwischen der Datenrezeption und der Ableitung von Schlussfolgerungen wurde für die Grafiken und Bezugsnormen untersucht, inwiefern Lehrkräfte, die bei der Rezeption etwa verstärkt auf die soziale Norm und/oder auf eine bestimmte Grafik eingehen, dies auch wieder bei ihren Schlussfolgerungen tun (Berechnung von ordinalen Korrelationen zwischen den Anteilen der Bezugsnormen und/oder Grafiken bei der Rezeption und den Schlussfolgerungen). Hier zeigten sich durchweg nicht-signifikante (frequentistische) Korrelationen. Bayes Faktoren für Kendall's τ mit *default Priors* (van Doorn et al., 2018) zeigten jeweils (eher schwache) Evidenz für die Nullhypothese, $\tau = 0$ ($.264 \leq BF_{10} \leq .561$), mit Ausnahmen der sozialen Bezugsnorm bei Grafik 1 ($BF_{10} = 1.445$) sowie der ipsativen Bezugsnorm bei Grafik 1 ($BF_{10} = 8.596$).

Forschungsfrage 3

Zur Beantwortung der dritten Forschungsfrage (Inwiefern zeigen datenkompetente Lehrpersonen komplexere Elaborationen bei der Rezeption ihrer Rückmeldungen?) wurden die einzelnen Anteile der höchsten und niedrigsten *graph literacy*-Stufe mit den Scores des Datenkompetenztests korreliert und angenommen, dass Personen, die im Datenkompetenztest besser abschneiden, größere Anteile an der höchsten und kleinere Anteile an der niedrigsten *graph literacy*-Stufe zeigen sollten. Beide ordinalen Korrelationen

waren nicht signifikant. Bayes Faktoren zeigten jeweils schwache Evidenz für die Nullhypothese (*reading beyond the data*: BF_{10} = .267; *reading the data*: BF_{10} = .499).

5.4.3. Zusammenfassung

Bezüglich der übergeordneten Fragestellung (Wie werden Rückmeldungen aus Vergleichsarbeiten rezipiert und inwiefern zeigen sich Zusammenhänge zwischen der Datenkompetenz (Test) und der Performanz (Think-Aloud-Protokolle)?) lassen sich die Ergebnisse der beiden Teilstudien folgendermaßen bündeln: In Teilstudie 1 zeigte sich, dass die Lehramtsstudierenden sehr häufig direkt in den Grafiken gegebene Informationen wie die Anteile der Klasse an einzelnen Kompetenzstufen entnahmen (*reading the data*). Kategorien von mittlerer bis hoher Komplexität (*reading between* und *reading beyond the data*) wie etwa die Aggregierung der Werte für einzelne Leitideen wurden deutlich weniger häufig und weniger häufig korrekt vergeben (vgl. Abbildung 7 und 8). Die Think-Aloud-Äußerungen in Teilstudie 2 wiesen in der Hauptsache mittlere Komplexität auf und zeigten eine starke Bevorzugung der sozialen Bezugsnorm, wobei die individuellen Verläufe für deutliche Unterschiede in den Rezeptionsprozessen der Lehrpersonen sprechen (vgl. Abbildung 9). Ca. die Hälfte der Lehrpersonen formulierte konkrete Implikationen, die andere Hälfte äußerte keine Schlussfolgerungen oder unspezifischen Handlungsbedarf. Die Analysen hinsichtlich der Konsistenz sprechen insgesamt eher gegen das Vorliegen einer Konsistenz zwischen der Datenrezeption und den Schlussfolgerungen hinsichtlich der Grafiken bzw. Bezugsnormen.

In Teilstudie 1 zeigten sich zwar signifikante Effekte für die Prädiktion des Erreichens höherer *graph literacy*-Stufen durch die Scores des Datenkompetenztests, allerdings nicht durchgängig. Die Analysen in Teilstudie 2 dagegen deuten eher auf einen Nullzusammenhang hin.

5.5. Diskussion

Um Verarbeitungsprozesse von Rückmeldedaten bei Lehrpersonen (Altrichter et al., 2016; Schildkamp, 2019) zu untersuchen, sind Think-Aloud-Studien aus Sicht der Autor*innen sehr gut geeignet, wenngleich bei beiden Teilstudien Limitationen zu berücksichtigen sind: Beide Stichproben sind Gelegenheitsstichproben. Teilstichprobe 2 ist zudem eher klein und heterogen, weshalb die statistische Power tendenziell klein ist und potenzielle Moderatoren (wie bspw. die wahrgenommene Bedeutsamkeit von Vergleichsarbeiten) unberücksichtigt bleiben; ferner können aufgrund unterschiedlicher Testleiter*innen Testleitungseffekte bei den Think-Aloud-Interviews nicht ausgeschlossen werden. Weiterhin erlaubt die Auswertung der Think-Aloud-Protokolle über *time samplings* in Studie 2 die Untersuchung kognitiver Prozesse und Aussagen über prozentuale zeitliche Anteile zur Komplexität der Rezeption und zur

Adressierung der Bezugsnormen. Allerdings bleibt zu diskutieren, inwiefern quantitative Anteile Aussagen über die Qualität der Datenrezeption und der Schlussfolgerungen insgesamt sowie insbesondere zur Frage nach der Konsistenz der Schlussfolgerungen aus der Rezeption ermöglichen. Künftige Forschungsvorhaben könnten die Frage der Konsistenz z.B. adressieren, indem die verbalisierten Schlussfolgerungen und die Datenrezeption jeweils inhaltsanalytisch kodiert und innerhalb der Lehrpersonen verglichen werden.

Mit der gebotenen Vorsicht kann jedoch festgehalten werden, dass die Studie durchaus Erkenntnisse bezüglich der Datenkompetenz von Lehrpersonen allgemein und der spezifischen Performanz bei der Datenrezeption im Rahmen datenbasierter Unterrichtsentwicklung bei Lehramtsstudierenden und Lehrpersonen ermöglicht. Die Analyse der Rezeptionsprozesse legt nahe, dass die Lehramtsstudierenden sowie die Lehrkräfte, die zwar kaum die Einzelergebnisdarstellungen für einzelne Schüler*innen nutzten, nur niedrige bis mittlere Aggregierungsprozesse mit den Daten zeigen. Damit konzentrieren sie sich eher auf die in den Rückmeldungen direkt gegebenen Entitäten (z.B. Anteil der Klasse mit Kompetenzstufe 2 oder Lösungshäufigkeiten der Klasse bei einzelnen Aufgaben). Dieses Ergebnis zeigt sich über beide Teilstudien hinweg und damit unabhängig von Kontextwissen und möglichem Erfahrungswissen aufgrund der Durchführung von Vergleichsarbeiten. Somit generieren beide Studien die Hypothese, dass (angehende) Lehrpersonen Schwierigkeiten haben, die Ergebnisse in Rückmeldungen aus Vergleichsarbeiten bei der Rezeption stärker zu aggregieren und mit Kontextinformationen in Verbindung zu setzen. Für die Gesamterfassung der grafisch repräsentierten Informationen, die Adressierung und das In-Beziehung-Setzen mehrerer Bezugsnormen sowie die Berücksichtigung von Kontextinformationen ist allerdings die höchste *graph literacy*-Stufe notwendig. Damit reiht sich auch die vorliegende Studie in die Befunde derjenigen (Selbstauskunfts-)Studien ein, die darauf hindeuten, dass Lehramtsstudierende und Lehrpersonen Schwierigkeiten beim eigentlichen Verstehen statistischer Daten (auch aus Vergleichsarbeiten) haben (Koch, 2011, 2013; Schliesing, 2017) und geringe Aggregierungsstufen präferieren (Groß Ophoff, 2013b; Maier et al., 2012), obwohl letztere die Darstellungen als verständlich einschätzen.

Als eine große Stärke von Vergleichsarbeiten gilt die Bereitstellung kompetenzorientierter Leistungsmessungen, die sich an kriterialen Maßstäben bzw. Bildungsstandards orientieren. Die dominierende Rezeption ihrer eigenen VERA-Ergebnisse im sozialen Vergleich bei den Lehrpersonen in Teilstudie 2 zeigt eine geringere Wahrnehmung dieser kriterialen Informationen, die aber für die Initiierung von Unterrichtsentwicklungsmaßnahmen bedeutsam scheint (Groß Ophoff, 2013a, 2013b). Eine (zusätzliche) ipsative Normierung, um Stärken und Schwächen der Klasse zu erfassen, scheint aus konzeptioneller Sicht sehr sinnvoll, zeigt sich

aber (mit Ausnahme einer Lehrperson) kaum in den Ergebnissen dieser Studie. Das Ergebnis, dass rund die Hälfte der Lehrpersonen konkrete Maßnahmen auf Basis der VERA-Ergebnisse ihrer Klassen konstruierte, während die andere Hälfte keinen oder unspezifischen Handlungsbedarf formulierte, könnte zunächst damit erklärt werden, dass die VERA-Ergebnisse für die Lehrpersonen nicht relevant (genug) waren oder für sie keine Schlussfolgerungen für die Unterrichtsentwicklung nahelegten und sie daher auch keinen Handlungsbedarf bzw. Maßnahmen äußerten. Es korrespondiert allerdings auch mit Ergebnissen anderer Studien u.a. zu Vergleichsarbeiten, die nahelegen, dass es für Lehrpersonen herausfordernd scheint, die Informationen zu den Leistungen ihrer Schüler*innen zu verwenden und in pädagogische Handlungen zu transformieren (Altrichter et al., 2016).

Bezüglich der Assoziation der allgemeinen Datenkompetenz mit der Performanz der Think-Aloud-Protokolle ergibt sich ein heterogenes Bild: Die Ergebnisse von Teilstudie 1 zeigen schwache Evidenz für einen positiven Zusammenhang, während die Ergebnisse in Teilstudie 2 eher auf einen Nullzusammenhang hindeuten. Daraus lässt sich die Hypothese ableiten, dass die generische Datenkompetenz bei Lehramtsstudierenden, die mit Rückmeldungen von Vergleichsarbeiten nicht vertraut sein dürften, eine größere Rolle bei der Performanz in der Rezeption spielt in dem Sinne, als dass es datenkompetenteren Lehramtsstudierenden leichter fällt, komplexere Elaborationen vorzunehmen, wenn sie mit den Grafiken nicht vertraut sind. Bei Lehrpersonen aus der Praxis, die über Kontextwissen und Erfahrungen verfügen, könnten hingegen andere Determinanten wie (individuell und organisational) etablierte Routinen, Motivation, *beliefs* u.ä. (Coburn & Turner, 2011) überwiegen.

Zukünftige Forschungsfragen, die sich auf Basis dieser explorativen Studie ergeben, betreffen zunächst die Frage nach dem Einfluss der allgemeinen Datenkompetenz (und möglicher Moderatorvariablen) auf die Rezeptionsprozesse in der Performanz auf Individualebene, wie er auch theoretisch (Coburn & Turner, 2011; Helmke & Hosenfeld, 2005) angenommen wird. Da sich Zusammenhänge zwischen der generischen Datenkompetenz und der Komplexität der Datenrezeption nur bei den Lehramtsstudierenden aber nicht bei den Lehrpersonen aus der Praxis zeigten, könnte hier insbesondere die Rolle von Kontextinformation (über die nur Lehrpersonen bzgl. der von ihnen unterrichteten Klassen verfügen) sowie Erfahrungen bzw. Routinen, Motivation, *beliefs* u.a. im Umgang mit Vergleichsarbeiten fokussiert werden. Dabei könnte außerdem experimentell untersucht werden, wie sich die Förderung der allgemeinen Datenkompetenz einerseits sowie eine spezifisch auf Vergleichsarbeiten ausgerichtete Förderung andererseits auf Rezeptions- und Interpretationsprozesse auswirken. Weiterhin liegen bereits wirksame Interventionsstudien zur Förderung verschiedener Aspekte der

Datenkompetenz sowohl bei Lehrpersonen (Ebbeler et al., 2017; Koch, 2013; van Geel et al., 2016) als auch bei Lehramtsstudierenden (Merk et al., 2020; Reeves & Honig, 2015) vor, deren Bedarf (z.B. hinsichtlich der Gesamterfassung grafisch repräsentierter Informationen) und Implementation in der Praxis auch die Ergebnisse dieser Studie unterstreichen. Vor allem die Adressierung solcher Fragen im Rahmen der Lehramtsausbildung, auch in Verbindung mit fachlichen und fachdidaktischen Fragen zur Förderung der Kompetenzorientierung und eines kriterialen (und ipsativen) Fokus, scheint sinnvoll, insbesondere, um auch die Ableitung konsistenter unterrichtlicher Maßnahmen anbahnen zu können. Denn damit die zentrale und gleichzeitig äußerst voraussetzungsvolle Idee datenbasierter Schul- und Unterrichtsentwicklung, die Förderung der Leistungen von Schüler*innen durch die Weiterentwicklung unterrichtlichen Handelns auf der Grundlage von (Leistungs-)Daten, gelingen kann, ist die adäquate Rezeption der rückgemeldeten Informationen zwar ein notwendiger aber keinesfalls hinreichender erster Schritt. Dieser bedarf sowohl weiterer Forschung als auch der Implementierung gezielter Unterstützung für Akteur*innen in der Praxis.

6. How do teachers make sense of technology-based formative assessments? Results from process mining of think-aloud data (Artikel 2)

Bez, S., Burkart, F., Tomasik, M. J., & Merk, S. (revise and resubmit). How do teachers make sense of technology-based formative assessments in their daily practice? Results from process mining of think-aloud data. *Learning and Instruction*.

Stichworte
think-aloud, technology-based formative assessment, sensemaking, teachers, data-based decision making, process mining

Abstract
Background: Technology-based formative assessments are considered promising in terms of teacher relief and validity advantages. However, in practice, teachers' sensemaking of the assessment results and their construction of instructional implications for adaptive teaching are very important for the successful use.
Aims: We explored how teachers make sense of technology-based formative assessment with a strong focus on ecological validity and a process perspective using think-aloud methodology.
Sample: Forty-eight teachers participated in the study.
Methods: We asked the teachers to verbalize their thoughts while they made sense of their students' formative assessment results as they usually do. Screencasts of the verbalizations and assessment results were recorded. Based on these, trained raters coded the main steps of sensemaking and identified which specific aspects of the results were noticed based on a deductive-inductive coding scheme. Cluster analyses were applied to explore differences among teachers, and process mining was conducted to explore the main processes.
Results: We found four main steps in sensemaking: *noticing results*, *comparing with personal perspective*, *analyzing errors* and *constructing instructional implications*. Relative durations of these steps vary substantially among teachers. Cluster analyses indicate that sensemaking processes were differentiated according to the complexity of summarizing and building relationships between single data points. The discovered process model revealed low dependency values in general and indicates that noticing results on its own seemed to be insufficient for constructing instructional implications.
Conclusions: This study generates the hypothesis that analyzing errors and comparing results with the personal perspective play an important role for constructing instructional implications.

Formative assessment is considered to support students' learning (e.g., Kingston & Nash, 2011; Lee et al., 2020; Xuan et al., 2022) and, in recent decades, has received increased attention in the context of both policy and practice around the world (Birenbaum et al., 2015). In the context of adaptive teaching, formative assessment is highlighted as essential because (ongoing) assessing students' individual characteristics and current understanding seem to be necessary prerequisites for setting adequate individual learning goals as well as providing tailored instruction and giving individual support (Corno, 2008; Hardy et al., 2019). Given the rise of digitization in education, technology-based formative assessments, in particular, are attributed some promising opportunities, especially related to validity, including task pools with

calibrated items and sophisticated psychometric models (e.g., McLaughlin & Yan, 2017; Spector et al., 2016). In recent years, technology-based formative assessment systems have been developed and implemented for various domains in different countries (e.g., The Netherlands, Faber et al., 2017; Faber & Visscher, 2018; Switzerland, Tomasik et al., 2018). At the same time, several current reviews point out that teachers play a key role in the successful use of formative assessments in the classroom (Heitink et al., 2016; Schildkamp et al., 2020; Yan et al., 2021), especially with regard to their adequate noticing of and interpretation of the assessment results: If teachers do not notice important aspects of the assessment results or misinterpret the information encoded in the data, one can conclude that setting adequate learning goals as well as providing tailored instruction and giving individual support will be inappropriate. From the perspective of consequential validity, providing valid formative assessments based on technology is not sufficient on its own, and adequate interpretations and inferences are essential (Kane, 2013). However, little is known about teachers' actual activities in terms of using (formative) assessment data in practice (Hebbecker et al., 2022; Mandinach & Schildkamp, 2021a), especially the ways teachers notice and interpret technology-based formative assessments and how they construct implications for adaptive teaching in their daily practice. Against this background, we investigated how teachers make sense of technology-based formative assessment results with a strong focus on ecological validity and a process perspective using think-aloud methodology.

In this paper, we first provide an overview of formative assessment, especially with regard to the benefits of technology-based approaches. We then outline why teachers' sensemaking of technology-based formative assessment results in daily practice is important and how concurrent think-aloud can provide valuable insights in addition to other precious methodological approaches. Subsequently, we present our research questions and describe the present study.

6.1. Teachers' sensemaking of technology-based formative assessment

6.1.1. Formative assessment

Formative assessment can be conceptualized as a process, in which students' learning is diagnosed with the goal of adapting future instruction and supporting students in achieving their learning goals, especially by providing feedback (Black & Wiliam, 2009). In contrast to summative assessment, often termed assessment *of* learning, which takes place at the end of a learning sequence, formative assessment is conducted during and for the learning process (Cronbach, 1964). This is highlighted by the (often) interchangeably used term assessment *for* learning (Bennett, 2011; Scriven, 1967). Although there are many different

approaches and methods, according to Black and Wiliam's definition, the central characteristic of formative assessment is "that evidence about student achievement is elicited, interpreted and used [...] to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions [...] in the absence of the evidence" (Black & Wiliam, 2009, p. 9). In general, formative assessment can be seen as part of data-based decision making in education (Schildkamp, 2019). Concerning the empirical evidence that formative assessment can foster learning, older meta-analytic findings that reported moderate to large positive effect sizes have been scrutinized in plausible ways (Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011), for example, with regard to several methodological issues. More recently published meta-analyses (Lee et al., 2020; Xuan et al., 2022) report substantial positive effect sizes on learning (Lee et al., 2020: d = 0.29; Xuan et al., 2022 [focus on reading]: d = 0.19, 95% CI = [.14, .22]) after adjusting for publication bias.

6.1.2. Benefits of technology in the context of formative assessment

As outlined in the previous section, substantive evidence indicates that formative assessment can foster instruction and learning. With the rise of technology, it is argued that technology can provide several benefits for formative assessment. First, it is emphasized that technology-based formative assessment can reduce the time and effort required to conduct formative assessments and therefore relieve teachers: For example, item banks that include various task types (Conole & Warburton, 2005; McLaughlin & Yan, 2017) can reduce the burden on teachers by saving them the time and effort needed to create assessments on their own. Moreover, such platforms can automatically correct students' answers (Conole & Warburton, 2005). Based on that, they can provide (more or less elaborated) feedback for all students and teachers (Spector et al., 2016) by visually displaying the assessment results of single students or the learning group as a whole. Second, it is emphasized that technology-based formative assessment systems can achieve a high level of validity as they can use large item banks with tasks of calibrated difficulty (Spector et al., 2016) and can integrate objective scoring rules during the analysis of students' answers (Tomasik et al., 2018). They can also rely on sophisticated psychometric models that enable adaptive testing, comparisons with other learners and the measurement of the learning progress over time at the individual level (Conole & Warburton, 2005). In particular, the valid, longitudinal measurement of the learning progress of all students is a promising opportunity because it hits the central pedagogical concept of formative assessment: The valid measurement of learning progress in comparison to learning goals enables teachers to adapt instruction continuously as they can evaluate feedback, instruction and learning goals based on the learning progress of individual students.

6.1.3. (Capturing) Teachers' sensemaking of assessment results

Given the advantages of technology-based formative assessment in terms of reducing teachers' workloads and increasing validity, along with the increasing availability of such systems (Faber et al., 2017; Faber & Visscher, 2018; Hall et al., 2015; Hebbecker et al., 2022; Tomasik et al., 2018), nevertheless, teachers play an important role when it comes to the successful use of these systems in classrooms. In particular, teachers' sensemaking of assessment results is crucial from both a conceptual and an empirical point of view, as is outlined in the next section.

In general, formative assessment is considered a part of data-based decision making (Schildkamp, 2019). Current conceptual models of data-based decision making (e.g., Coburn & Turner, 2011; Mandinach & Gummer, 2016; Marsh, 2012; Schildkamp, 2019) differ in their details but they all conceptualize a sequential process with several steps that follow each other (Bez et al., 2023; Hebbecker et al., 2022): First, data must be collected. Then, teachers have to make sense of the data, which means that they have to notice relevant aspects of the data and interpret them in order to formulate corresponding instructional implications. These implications have to be put into practice and evaluated in their outcomes. Since sensemaking is the first step after data collection that can be simplified and improved through technology-based systems, one can conclude that appropriate sensemaking is crucial for the following steps. Similarly, from the perspective of consequential validity (Kane, 2013), one can argue that noticing and interpreting technology-based formative assessment results is a necessary prerequisite for making adequate inferences in terms of constructing pedagogical implications and instructional decisions. If teachers do not notice important aspects in assessment results that are provided with high construct validity by a technology-based assessment system or if they misinterpret information, the derived instructional conclusions and decisions as well as the following specific instructional actions will be necessarily inadequate and therefore consequential validity is diminished.

From an empirical perspective, several systematic reviews have emphasized the importance of teachers, especially their data literacy, for successful formative assessment in practice (Heitink et al., 2016; Schildkamp et al., 2020; Yan et al., 2021). Research focusing on teachers' use of data indicates that teachers often find this challenging (Mandinach & Schildkamp, 2021a) and generally tend to have difficulties making sense of data and deriving conclusions for instructional decisions (e.g., Kippers et al., 2018; van den Bosch et al., 2017; Zeuch et al., 2017). However, little previous research has provided insights into teachers' actual activities regarding (formative) assessment data in practice (Hebbecker et al., 2022; Mandinach & Schildkamp, 2021a), especially at the micro level of sensemaking (Bez et al., 2021; Goffin et

al., 2022). To understand teachers' actual practices, retrospective self-reports are limited because they can be biased and are restricted in their opportunity to provide insights into micro processes in vivo. Test scores (e.g. from data literacy tests) do not necessarily have to correspond with the performance of teachers in practice, and inferences based on intervention studies with a focus on internal validity (e.g. on data literacy) have (necessarily) limitations related to ecological validity. Log-file analyses might be promising in investigations of technology-based formative assessments as they describe actual behavior objectively. However, the construct validity of measures such as dwell time or click sums remains difficult (Hebbecker et al., 2022) as their interpretations are often ambiguous. That means that inferences based on log data about the cognitions of teachers, in terms of what they notice in assessment results and how they interpret the information and construct instructional implications, are limited. In this context, think-aloud is considered an appropriate methodological approach (Espin et al., 2017), because it can provide non-reactive insights into everyday life cognitive processes (Ericsson & Simon, 1998; Fox et al., 2011). In particular, concurrent think-aloud seems adequate because, in contrast to retrospective think-aloud, it addresses processes in the working memory (Leighton, 2017). Therefore, concurrent verbalizations of thoughts provide highly objective (i.e., minimally intrusive) and ecologically valid (i.e., related to daily life) insights into cognitions (i.e., a process perspective).

Indeed, some think-aloud studies addressing teachers' reading and interpreting of (formative) assessment data (Espin et al., 2017; Goffin et al., 2023; van den Bosch et al., 2017; Wagner et al., 2017) have already been conducted: However, only Goffin et al. (2023) and van den Bosch et al. (2017) used the assessment results of the students of the participating teachers. Regarding the think-aloud approach, in all of the mentioned studies the participants were asked to think aloud and explain their thoughts, too. According to Ericsson and Simon (1998) and Fox (2011), this mode of think-aloud must be differentiated from a purely spontaneous open think-aloud because, in the first mode, the participants' thoughts are affected by the elicited descriptions and explanations.

6.1.4. The present study

As outlined in the previous sections, technology-based formative assessments are promising in terms of their ability to relieve teachers and their advantages in terms of validity. However, a crucial prerequisite for the success in practice is teachers' sensemaking of assessment results and their construction of instructional implications for adaptive teaching. However, little is known about how teachers make sense of technology-based formative assessments in their daily practice. Against this background, our research goal is to gain insights into teachers'

sensemaking with a strong focus on ecological validity and a process perspective using think-aloud methodology. Thus, our research questions are as follows:

- Research Question 1 (RQ 1): Which addressed aspects and process steps show think-aloud data obtained from teachers as they make sense of technology-based formative assessments?
- Research Question 2 (RQ 2): Can different groups of teachers be identified according to which aspects of formative assessment results are addressed and which process steps are shown?
- Research Question 3 (RQ 3): Which typical sensemaking processes in teachers' think-aloud data can be identified?

6.2. Method

6.2.1. Participants

We conducted think-aloud sessions with $N = 48$ in-service teachers ($M_{age} = 44.1$, $SD_{age} = 11.6$, $M_{expertise} = 16.5$, $SD_{expertise} = 10.5$, 56% identified themselves as male). All participating teachers are located in Switzerland and are voluntary users of the technology-based formative assessment platform Mindsteps (www.mindsteps.ch). Of the teachers, 31% work as primary school teachers, 69% work as secondary school teachers, and 90% teach at least one STEM subject (94% teach at least one language, 80% teach a social science subject, and 50% teach art). Participants were invited to participate via email.

6.2.2. Design

To investigate how teachers make sense of technology-based formative assessment in their daily practice, we conducted an exploratory study with the focus on ecological validity and a process perspective using concurrent think-aloud methodology. Therefore, we included participants who use a technology-based formative assessment platform (www.mindsteps.ch) in their daily practice voluntarily. We decided to conduct think-aloud sessions with those teachers focusing on their students' latest formative assessment results and not on artificial assessments or assessment results of other students etc. Executing concurrent think-aloud, the participants were asked to verbalize their thoughts while they made sense of students' assessment results as they usually do (for details, please see Section 6.2.3). To capture cognitive sensemaking processes with a focus on ecological validity in terms of daily behavior and to prevent biases and reactivity, we avoided making directed requests (e.g., for explanations, descriptions) because they are considered to affect performance and sequences (Fox et al., 2011; Leighton, 2017). To reduce the time and effort required of the participants and to get in touch with them in their daily environment, the sessions were

conducted via video calls. During the sessions, the participants logged into their formative assessment platform account and shared their screens. The participants' screens as well as their verbalizations were recorded. Trained raters coded these audiovisual data (verbalizations and screencasts) using an inductive-deductive developed coding scheme (see Section 6.2.3). The resulting data formed the basis for further statistical analyses (see Sections 6.2.3 and 6.2.4 for details). The ethics committee of the respective university approved the study.

6.2.3. Data collection and coding procedure

The one-to-one think-aloud sessions were conducted by two trained interviewers and were structured based on the guidelines given in a detailed manual (a translated version can be found on the Open Science Framework [OSF], https://bit.ly/osf01). The sessions were divided into several parts: After a short check of the technical setup, participants were introduced to the think-aloud procedure, and a short warm-up was conducted, as recommended in the literature (e.g., Leighton, 2017). After that, the concurrent think-aloud phase of the session was prompted as follows: "Please express everything that you are thinking while you are looking at the current assessment results of your class the way you usually do and point with your mouse at the corresponding areas. When you have finished, please say 'I have finished.'" In this phase, cameras of both the teacher and the interviewer were turned off in order to avoid unintended affections or distractions through facial expressions. The interviewer strictly stayed in the background and only said "please keep talking" or "please use your mouse", if necessary (Padilla & Leighton, 2017). When the teacher was finished, the interviewer asked the participant what the key information was in the results, whether they derived any conclusions from the assessment data and whether they would take specific instructional or general actions based on the assessments. The current paper focuses on the data obtained from the concurrent think-aloud phase with the participants. During the think-aloud phase, the participants clicked through the platform and looked at different graphs encoding the assessment results of their students. Appendix A provides exemplary screenshots of the different graph types.

To to provide insights into the sensemaking processes on the micro level, in the first run of coding, we maintained the sequential structure of the think-aloud data; therefore, the audiovisual data (recorded screencasts and verbalizations) were coded as timed-event codings (Bakeman & Quera, 2011), accurate to the second using qualitative data analysis software. The coding scheme was inductive-deductive, developed based on the framework provided by Coburn and Turner (2011). It was captured which graph type and, independently of this, which sensemaking step was addressed. Two trained raters coded independently from

each other and then discussed any disagreements until they reached a consensus. Due to a sufficient coding agreement of Krippendorff's $\alpha \geq .6$ (Hayes & Krippendorff, 2007) and the need for economic efficiency, 60% of the think-aloud sessions were coded using this procedure, and 40% were coded by one rater. We captured four main steps in the teachers' sensemaking: *noting results*, *comparing with the teacher's personal perspective*, *analyzing errors* and *constructing instructional implications*. During the coding process, we faced the known issue of co-occurrence and exclusivity of codes (Bakeman & Quera, 2011): In some cases, two sensemaking steps, especially *noticing results* and *analyzing errors*, were closely entangled, so we followed Bakeman and Quera (2011) and determined these cases later in the analysis.

To capture which aspects in the rich and extensive assessment results provided by the assessment platform were noticed, we conducted a second run of coding. It was captured based on an inductive-deductive developed coding scheme, whether specific aspects given in the displayed assessment results were noticed at least once, e.g., mean score of class or groups of students according to the results. This run of coding covered only think-aloud data related to graph 2 (see Appendix A) because this graph type was the one that was addressed mostly by the teachers (see Appendix B). We used Cohen's *k* to calculate coding agreement and percentage agreement if *k* could not be calculated, occurring when one rater never assigned a code. Due to sufficient coding agreement ($.5 \leq k \leq 1$; outlier $k \leq .4$; percentage agreement ≥ 83%), considering that Cohen's *k* is reduced in the case of low prevalence of codes (Bakeman & Quera, 2011), and the need for economic efficiency, again 60% of the think-aloud sessions were coded independently by two trained raters, who discussed any disagreements until consensus was reached, and 40% were coded by one rater. The coding scheme (translated and shortened) is provided on OSF (https://bit.ly/osf01).

6.2.4. Data analysis

Our research questions focus on teachers' sensemaking of formative assessment data. To investigate which process steps can be found in the think-aloud data and how prevalent they are (RQ 1), we first preprocessed the timed-event codings to time samplings (Bakeman & Quera, 2011) with the interval of one second, calculated descriptive statistics and visualized the codings in their sequences for each teacher.

To explore the patterns of different groups of teachers in terms of the relative durations of the main steps and the aspects they addressed (RQ 2), we applied cluster analyses. As the data basis contained both continuous variables (e.g. proportions of main steps) and binary variables (e.g. noticed aspects), we decided to use Gower's coefficient (Gower, 1971) as

similarity measurement (Kaufman & Rousseeuw, 2005), and we chose the average and complete linkage algorithms for clustering. Additionally to a visual inspection of the results in heatmaps with dendrograms, partitions of clustering solutions were compared using the Adjusted Rand Index (Hubert & Arabie, 1985) to aim for an appropriate and robust result.

To explore the think-aloud data (based on the codings) from a process perspective (RQ 3), we applied process mining because it can capture a typical process in event data and describe it in a sequential model (Reimann, 2009). Due to the exploratory design of the study and the absence of a preexisting process model, we decided to use discovery and not conformance checking (van der Aalst, 2016) and applied the Flexible Heuristics Miner Algorithm (Weijters & Ribeiro, 2011). This algorithm is appropriate for think-aloud data in educational research (e.g., Hartmann et al., 2022; Sonnenberg & Bannert, 2019) since it discovers dependencies in the ordering of events, takes loops into account, and, in contrast to other process mining algorithms, is able to deal with low-structured or noisy data (van der Aalst, 2016; Weijters & Ribeiro, 2011). The basic concept behind this algorithm is to retrieve a causal process model based on dependencies between events using frequency-based metrics (Weijters & Ribeiro, 2011). That is, if event A is very often directly followed by event B in the data, but event B is very rarely directly followed by event A, the value of the dependency measure for "event A followed by event B" is high. This indicates that a causal dependency between event A and event B (in this order) is likely. The values of the dependency measures are always between -1 (low dependency) and 1 (high dependency). To discover the main process in teachers' sensemaking, we explored a dependency graph based on a dependency matrix using the heuristicsmineR package (Mannhardt & Janssenswillen, 2023). Due to the very limited knowledge about how to set various parameters (a priori), especially in low-structured domains, we followed Weijters and Ribeiro (2011) and specified a complete model with low frequency and dependence thresholds and a simplified model with high(er) thresholds. The reproducible documentation of the complete data analysis is provided on OSF (https://bit.ly/osf01).

6.3. Results

6.3.1. Research question 1

We found four main process steps in the teachers' sensemaking, as follows: *Noticing results* refers to the teachers' verbalizations when they notice the formative assessment results provided by the system, for example, "Tom got 546 points on the scale" or "mean of 477 points is quite good and concerning dispersion I see they are roughly all between 250 and 850". *Comparing system results with the teacher's personal perspective* covers the verbalizations of teachers when they compare the system's results or specific aspects with their own point

of view, for example, “I am surprised because Hannah is normally one of the best in class”, or “this is exactly what I expected”. *Analyzing errors* captures expressions when teachers analyze errors or try to elaborate on the mistakes students make or their misconceptions, for example, “funny, he has an error at bar charts, probably a comprehension problem” or “the other addition tasks were correct, so the problem here was this specific task type”. *Constructing instructional implications* covers teachers’ verbalizations when they construct conclusions for general or instructional future actions, for example, “many students had problems with these tasks so I will repeat this topic in the next lesson" or “I will talk to the student about the results”. The relative durations of these process steps during think-aloud are provided in Abbildung 10. All the teachers noticed the results but not all of the participants analyzed errors, made comparisons with their own perspectives or formulated instructional implications. The relative durations of the process steps show obvious variance among teachers, especially regarding *noticing results* and *analyzing errors*.

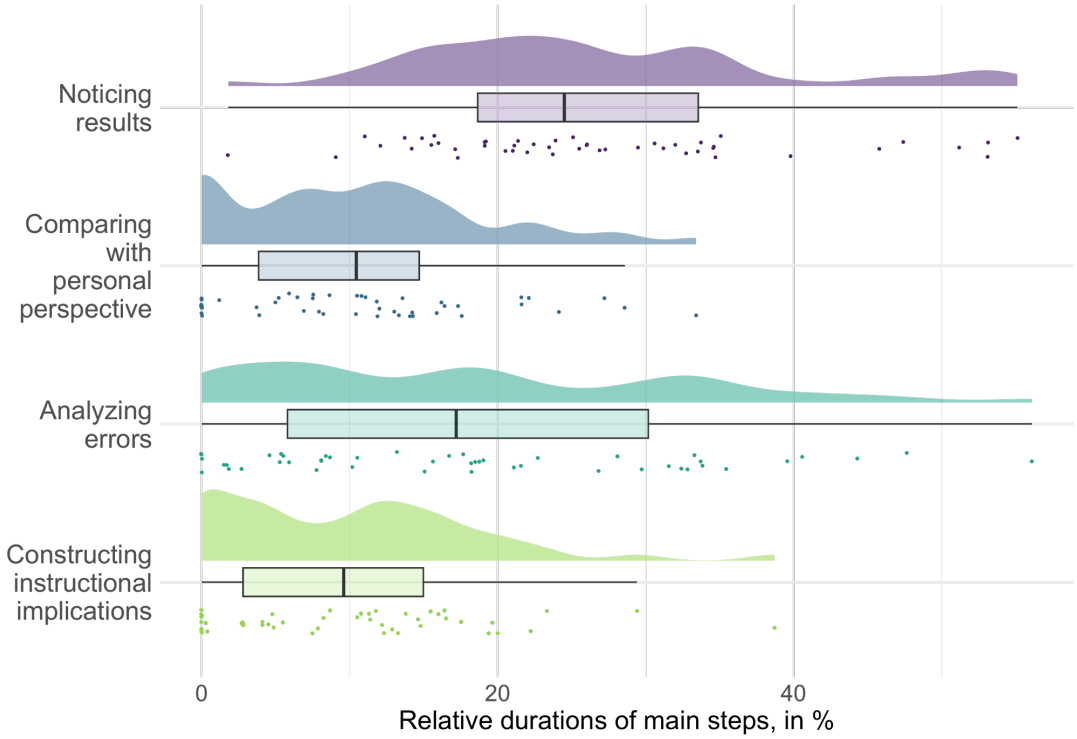


Abbildung 10. Relative durations of process steps during think-aloud

Notes. *Comparing with personal perspective* is an abbreviation for *Comparing results with the teacher’s personal perspective*. The Co-occurrences of main steps and the relative durations of addressing different graph types can be found in Appendix B.

Abbildung 11 shows the percentages of teachers who noticed specific aspects of the results presented in Graph 2 of the formative assessment system (see Appendix A, class overview) at least once. Results show that the highest percentages belong to those aspects of the results

which can be directly noticed without complex cognitive elaborations (e.g., matching results to students' names, contrary to elaborating dispersion in the results according to the average distance of data points from the mean). Thirty percent of the teachers grouped the best and about 25% of the teachers grouped the weakest students, but only 9% grouped the middle of the displayed results. Concerning verbalizations that address common concepts within descriptive statistics, the highest percentage addressed the mean (about 55%) whereas only about 10% and fewer of the teachers mentioned range, outliers or dispersion based on the average distance from the mean.

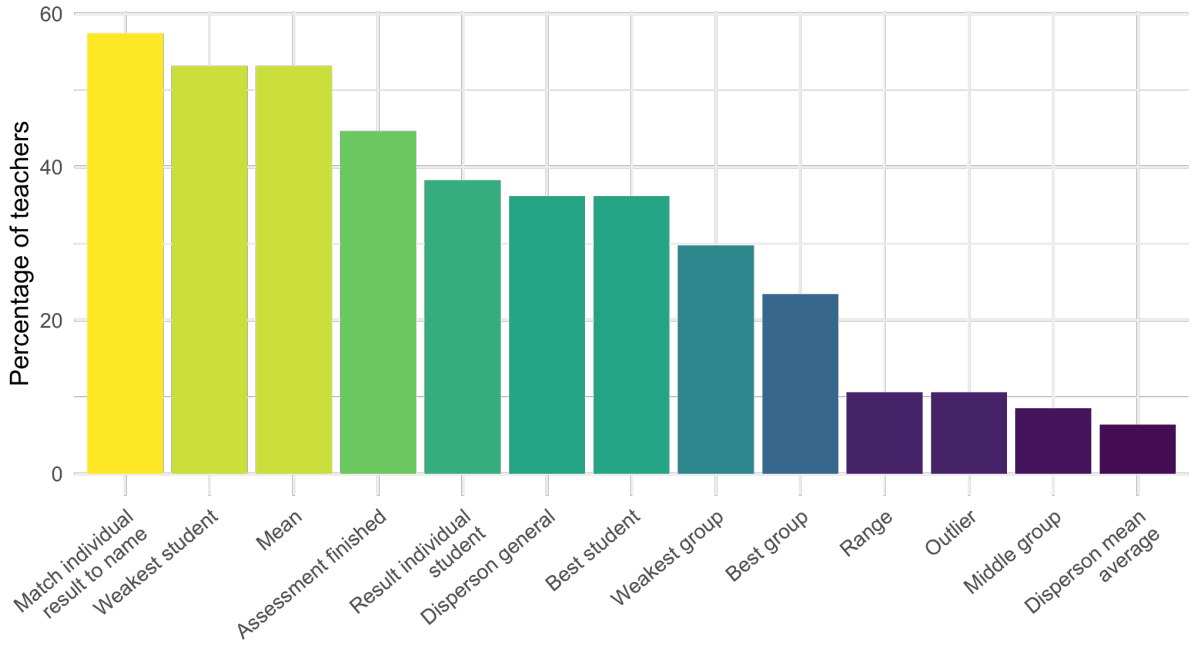


Abbildung 11. Percentage of teachers who noticed specific aspects of the results

Notes. The percentage of teachers who verbalized specific aspects of the results presented in graph 2 (class overview, see Appendix A) at least once in think-aloud.

To get a first impression of processes of sensemaking during think-aloud, we visualized the sequence of the main steps (respectively, the codings) for each teacher (Abbildung 12). This graphical overview shows substantial variance among teachers in terms of the sequences. For example, the visualized think-aloud data of several teachers show many recursions and iterations in their sequences (e.g., Tara and Van).

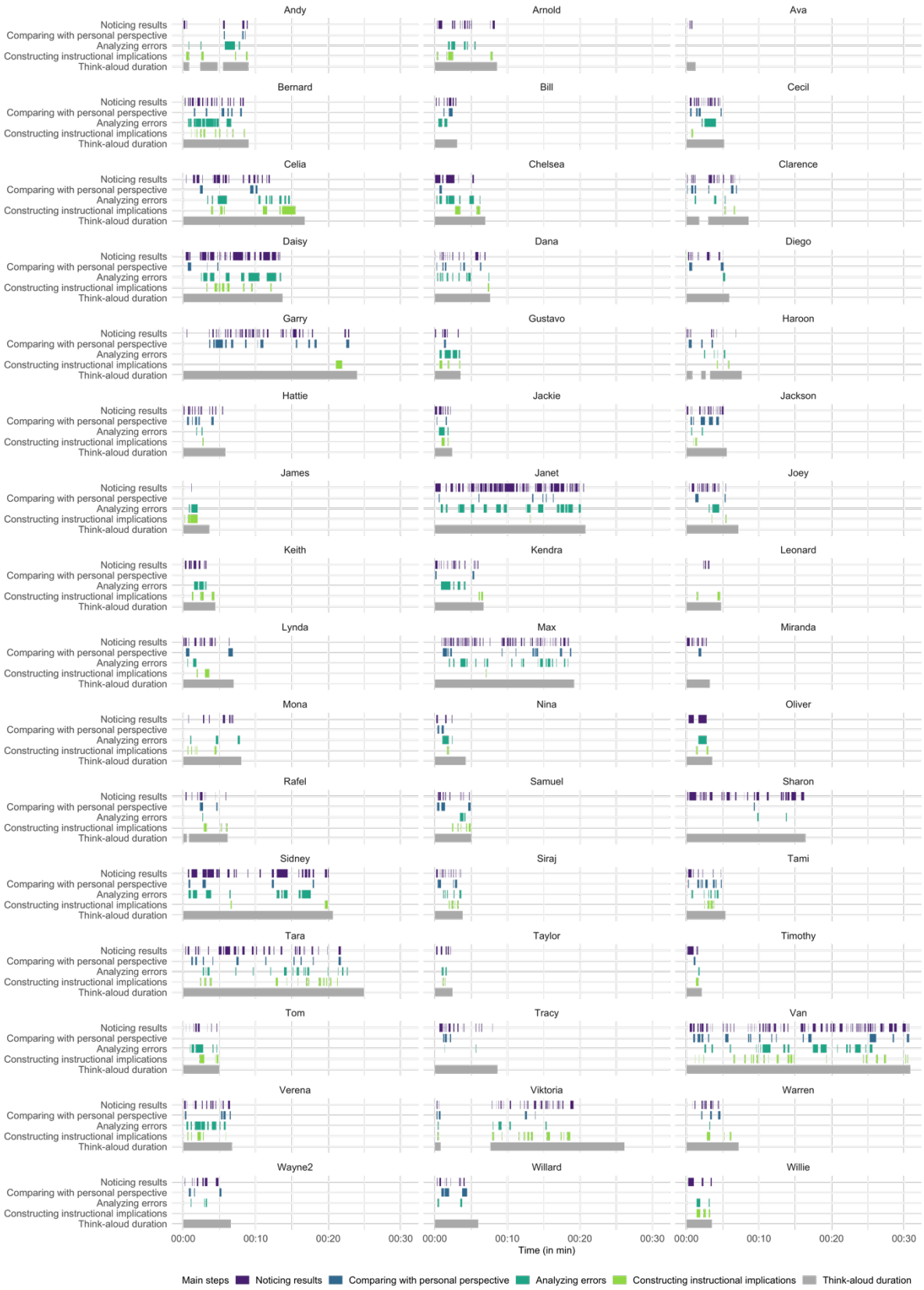


Abbildung 12. Sequences of main steps per teacher

Notes. A larger version of this figure with better resolution can be found on the OSF, https://bit.ly/osf01.

6.3.2. Research question 2

Beyond the descriptive statistics of sensemaking (RQ 1), we investigated whether the teachers could be differentiated into groups according to their sensemaking processes. To this end, we conducted cluster analyses based on the relative durations of the main process steps and the aspects of the results that teachers addressed in their sensemaking using average and complete linkage algorithms (based on Gower's distance matrix coefficient, [Gower, 1971]) and visualized the results using a heatmap with dendrograms (see Abbildung 13).

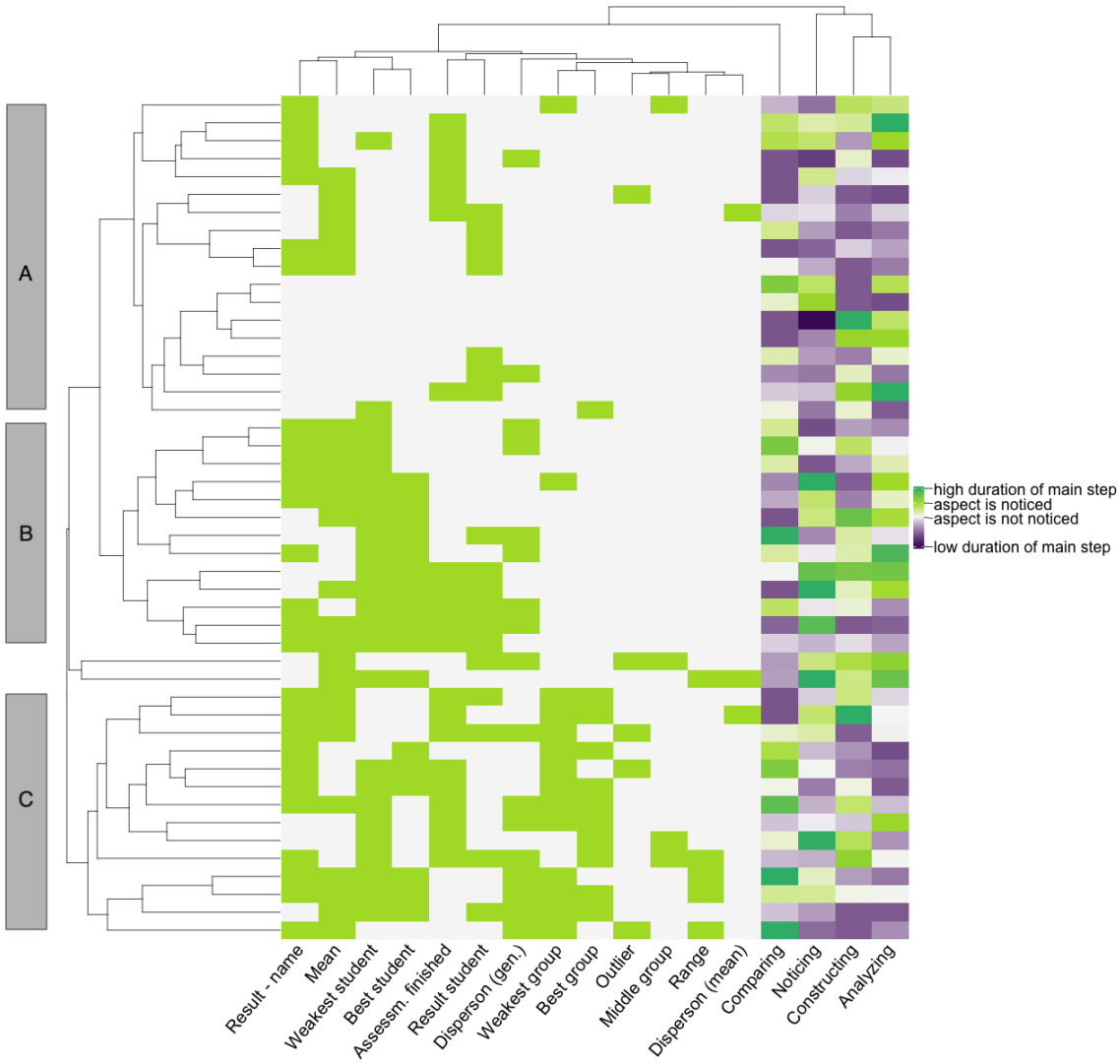


Abbildung 13. Visualization of cluster analysis

Notes. Heatmap and dendrograms using the average linkage algorithm based on Gower's similarity measure. Each row represents the data of one teacher. The gray boxes with letters on the left visualize groups (group A at the top, group B in the middle, and group C at the bottom).

According to Abbildung 13, it is plausible to assume that there are three main groups (group A, B, C) of teachers in the data. This partition, based on average linkage, was compared to the three-group solution of complete linkage to aim for a robust result. The Adjusted Rand

Index of $r = .9$ (values can theoretically range from -1 to 1) indicates a very high similarity between these two partitions (Hubert & Arabie, 1985). Group A (43% of participants, depicted at the top of the heatmap) notices no or only a few aspects that are directly given in the results of individual students and showed relatively low relative durations for each of the sensemaking steps. In contrast, group B noticed the best and the weakest student, and group C formed groups of students according to the best and weakest results. Differences in cognitions can be characterized mainly according to the complexity of summarizing and building relationships between single and directly given data points (e.g., noticing the best/weakest students vs. forming corresponding groups). This is plausible according to the concept of graph literacy (Friel et al., 2001; Galesic & Garcia-Retamero, 2011). It differentiates between levels of complexity in reading and comprehending graphs based on building relationships and summarizing data points as well as on making inferences on data: Reading data (the lowest level) covers the extraction of directly given entities in graphs, e.g., noticing means or finished assessments. Reading between data (middle level) captures building relationships or summarizing data points, e.g. forming groups. Reading beyond data (the highest level) means summarizing the graph as a whole and making inferences. Noticing of group A and B can be considered to reflect a rather low level of complexity, whereas the noticing cognitions of group C reflects mainly a middle level of graph literacy.

6.3.3. Research question 3

To explore the main processes in teachers' sensemaking of formative assessment results, we mined a dependency graph based on a dependency matrix using process mining (see Abbildung 14). We followed Weijters and Ribero (2011) and specified a complete model (including low-reliable dependency relations) and a simplified model (including only high-reliable dependency relations). The complete model (see Abbildung 14), including low thresholds on frequency and performance, is considered more appropriate to the data in this study, because higher default thresholds on dependency may lead to an oversimplified model (see reproducible documentation of data analysis for details, https://bit.ly/osf01).

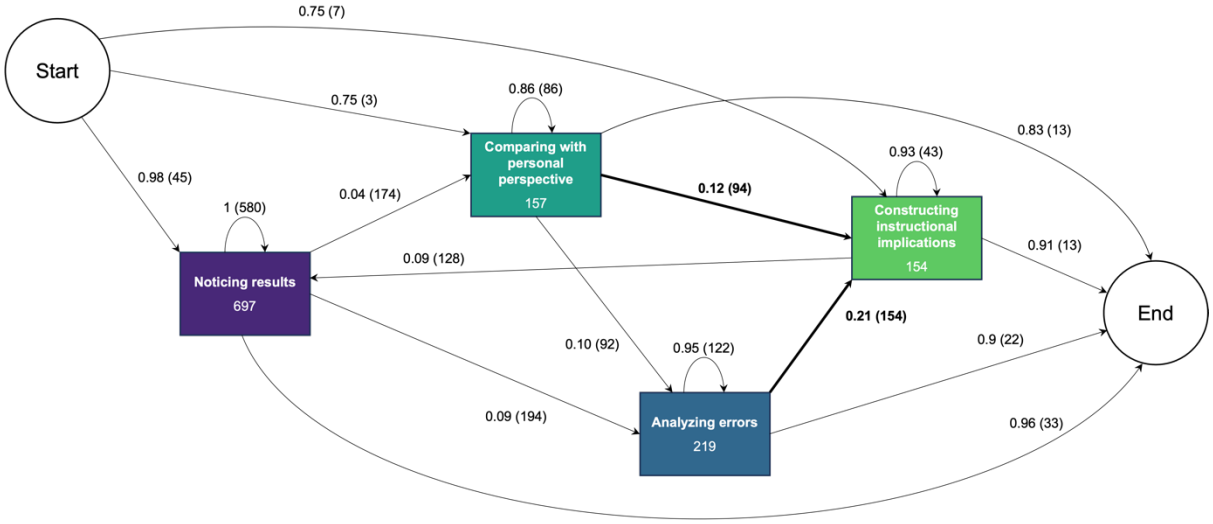


Abbildung 14. Process model of sensemaking

Notes. The process model represented as a dependency graph including frequencies. The boxes represent event classes (sensemaking steps, such as *noticing results*), and the numbers in the boxes indicate the frequency of the step's occurrence. The arcs show the dependencies between the steps. The left numbers near the arcs are dependency measures (indicating the strength of the dependency, between -1, low, and 1, strong); the right numbers (in brackets) display the frequencies of the transitions.

Generally, the model shows low dependency measures, except for length-one-loops. The latter indicates that codings of process steps tend to be followed by another coding event from the same process step. Although dependency measures from the antecedents *analyzing errors* and *comparing the results with the teacher's personal perspective* to the consequent *constructing instructional implications* are low, the model shows high frequencies for the corresponding arcs and no arc from *noticing results* to *constructing instructional implications*. This can be interpreted as an indication that strictly noticing of assessment results is not sufficient for constructing instructional implications; rather, the analysis of errors in the results and the comparison of the assessment results with the personal perspective are important steps between *noticing results* and *constructing instructional implications*. The process model indicates a sensemaking process in which information derived from formative assessment results is enriched by the teachers' own impressions as a form of contextual information and knowledge as well as by elaborations of students' mistakes, e.g. misconceptions or motivational problems that may lead to a specific pattern of student responses, and, based on that, conclusions concerning instructional implications and actions are derived.

6.4. Discussion

6.4.1. Summary of findings and conclusions

In this study, we explored how teachers make sense of technology-based formative assessments, with a strong focus on ecological validity and a process perspective using concurrent think-aloud. Concerning RQ 1, we found four main steps in the teachers'

sensemaking: *noticing results*, *comparing the results with the teacher's personal perspective*, *analyzing errors,* and *constructing instructional implications*. The relative durations of these steps in the think-aloud data varied substantially among the teachers. The step *comparing the results with the teacher's personal perspective* can be related to research focusing on beliefs and intuitions in data-based decision making of teachers, which has shown that data-based decision making can be affected by biases (e.g., Mandinach & Schildkamp, 2021a; Vanlommel et al., 2017). This step also corresponds to the model developed by Mandinach and Gummer (2016), in which they highlight that data use for teaching means that data are combined, understood, and integrated with different domains of professional knowledge, including knowledge of educational contexts and the characteristics of students, as well as content, curriculum and pedagogical (content) knowledge. The latter corresponds to the sensemaking step *analyzing errors*, which covers elaborations of students' mistakes involving corresponding domains of professional knowledge. To illustrate this sensemaking process, we give the following example on a conceptual level: A teacher notices that, on average, her class correctly solved 60% of the tasks provided by the technology-based formative assessment system (*noticing results*). She elaborates on what this result means in terms of average knowledge and competencies of the students, their misconceptions, patterns of errors, etc. (*analyzing errors*). She considers this in relation to her own perspective (*comparing the results with the teacher's personal perspective*), for example, challenges the students experienced in past lessons, her students' characteristics, and the assessment situation, and formulates future instructional implications (*constructing instructional implications*), for example, revising prior knowledge in the class, addressing specific misconceptions, or adapting learning goals for specific students.

In this study, teachers addressed many different aspects in the exhaustive assessment results provided by the system. Specific aspects that were verbalized by most of the teachers at least once have in common that they can be directly noticed without complex cognitive elaborations. This corresponds with the results of other studies focusing on the graph literacy (Friel et al., 2001; Galesic & Garcia-Retamero, 2011), which showed low to middle levels of performance of teachers (e.g., Bez et al., 2021; Zeuch et al., 2017). Regarding RQ 2, cluster analyses using different algorithms indicate three groups of teachers who differed in terms of the complexity they displayed in summarizing and building relationships between single data points, which is consistent with the concept of graph literacy. Concerning RQ 3, we discovered a process model by applying the Flexible Heuristics Miner. The complete model (based on low thresholds for frequency and performance) reveals low dependency values in general, except for length-one-loops. Due to the antecedents *comparing the results with the teacher's personal perspective* and *analyzing errors* to the consequent *constructing instructional implications*, one

can derive as a hypothesis that these steps play an important role for the step *constructing instructional implications*.

6.4.2. Limitations

Several limitations of this study arise due to the exploratory nature of the study and a potentially positive selected sample, which included only teachers who voluntarily use a specific technology-based formative assessment system. These limitations limit the generalizability of the findings, and all generated hypotheses need further confirmatory investigation in other contexts, for example, with teachers using different formative assessment platforms. Pivotal limitations of the study related to the think-aloud methodology lie in the prerequisite that it is not assumed that all thoughts of participants are captured (Fox et al., 2011). Furthermore, analyses based on a timed-event coding procedure in think-aloud must consider that think-aloud affects time (Fox et al., 2011), so time and durations have to be interpreted relatively. Due to this focus, we did not investigate the consistency or the plausibility of sensemaking, in terms of the quality of derived instructional implications based on the previous noticing results, the error analysis or the teachers' comparisons with their personal perspectives. Concerning our cluster analyses, we aimed to achieve robust and validated results by using different algorithms and indices to compare different partitions, in addition to visual inspections and theoretical considerations. However, obtaining appropriate cluster analyses remains challenging in general (Kettenring, 2006). Regarding discovering a process model using process mining, one has to consider the low values in dependency measures in the resulting model. This could be due to noisy data or low-quality data coding or no causal dependencies of the corresponding sensemaking steps in the cognitions of the teachers.

6.4.3. Practical implications and further research

Conceptual models of data-based decision making consist of a sequential structure and differentiate noticing data, interpreting and transforming information to instructional decisions (see Chapter 6.1.3). This corresponds to the sensemaking steps and their sequences that were found in this study. In addition to the current conceptual models, the sensemaking cognitions of many teachers show many recursions and iterations in their sequences (see Abbildung 12). If further research confirms this, then adding corresponding loops in the models may be reasonable. As already implied in previous sections, further research should focus on sensemaking processes relating to the consistency and plausibility of the sensemaking processes as well as the interplay between domains of professional knowledge and sensemaking. Furthermore, insights into user cognitions in daily practice can provide valuable information and implications for designing and improving technology-based formative

assessment platforms for teachers. For example, teachers may find having to click through several individual task results per student to find patterns of mistakes, such as related to specific misconceptions, to be cognitively exhausting and time consuming, and technology could better facilitate this. We propose that assessment tasks be developed and implemented that are very good at diagnosing students' misunderstandings or general difficulties in different areas. Based on that, technology could cluster groups of students according to their results and groups of tasks according to competence levels respectively knowledge domains. This may relieve teachers' sensemaking for adaptive instruction because it would allow them to more easily capture which students are having difficulties in which domains. Moreover, the results of this study indicate that cognitions of teachers differ in terms of the complexity displayed in summarizing and building relationships between data points in the assessment results. This could be taken as a starting point for developing adaptive support and training for teachers.

In this final section, methodological reflections and recommendations are outlined and discussed. Generally, we would argue that concurrent think-aloud is a valuable methodological approach for investigating cognitions in technology-rich settings with a focus on ecological validity. In our study, we conducted think-aloud sessions via video calls rather than in a laboratory setting with the aim of reducing the time and effort required from the participants and reaching them in their daily environment. However, further investigation is needed to understand the potential differences in remote think-aloud sessions versus face-to-face think-aloud sessions. Concerning the coding of think-aloud data, it is evident that the chosen coding approach determines further analyses: In some cases, “plain coding” without taking time into consideration and analyzing the resulting frequencies can be sufficient. The timed-event coding of videotaped (or audiotaped) think-aloud sessions requires more resources and time and can be challenging in terms of achieving adequate reliability, in our experience. However, the timed-event coding approach in combination with new methodological approaches and developments in data science, for example, process mining as it is applied in this study, enables the maintenance of a sequential structure and to investigate sequential processes in cognition in educational contexts. Process mining can be used for discovery, as we did in this study, but also conformance checking and enhancement as other types of process mining (van der Aalst, 2016) are worth considering in future research in learning and instruction.

7. Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers (Artikel 3)

Wurster, S., **Bez, S.**, & Merk, S. (2023). Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers. *Learning and Instruction*, *88*, 101806. https://doi.org/10.1016/j.learninstruc.2023.101806

Stichworte
Data-based decision making, Pre-service teachers, Data literacy, Motivational beliefs about data use

Abstract
Background: In times of expanding datafication in education, data-based decision making (DBDM) has become a crucial part of teaching and data literacy and motivational beliefs about data use are important preconditions of teachers' data use. However, data use is not yet part of most teacher education programs. Aims: The present study investigated the effects of a short and easily implementable online intervention on preservice teachers' data literacy and their motivational beliefs about data use.
Samples: We conducted two studies with pre-service teachers: in the pilot (Study 1), the structure of motivational beliefs about DBDM was investigated (N = 34). In Study 2 (the main study), the effects of the intervention were examined (N = 136).
Methods: The study was designed as a randomized controlled trial with a wait list control group. The intervention focused on building the basic concept of data literacy with emphasis on data analysis and different types of data, such as assessment and self-evaluation data.
Results: Study 1 revealed a three-factor structure of motivational beliefs about DBDM. The intervention in study 2 showed a strong positive effect on data literacy test scores and self-efficacy. On average, pre-service teachers reported positive motivational beliefs about DBDM. However, we found evidence for null effects regarding changes in motivational beliefs in the value components cost and enjoyment.
Conclusions: The intervention has the potential to foster aspects of data literacy and self-efficacy about DBDM but not value components of beliefs about DBDM in general.

Teachers around the world are expected to use data as a basis for their professional decisions on learning and instruction (Mandinach & Gummer, 2013; Mandinach & Schildkamp, 2021b; Schildkamp et al., 2013). Data has the potential to inform pedagogical decisions and thus affect instruction, especially if the notion of data includes not only test results but also informal data like data about the quality of instruction, behavioural data and data about absenteeism. However, data itself is not sufficient to initiate improvement (Hoogland et al., 2016; Schildkamp et al., 2017). Therefore, data has to be analysed and interpreted in its specific context, and further implications have to be constructed and translated into practice. Previous research shows mixed results regarding the effects of data-based decision making (DBDM) on learners

in terms of, for example, student achievement (Ansyari et al., 2020; Visscher, 2021). In general, DBDM can be situated between the poles of accountability and improvement (Mandinach & Schildkamp, 2021a). Stressing accountability purposes and a concurrent overreliance on achievement data can lead to side effects like “teaching to the test” (Hamilton et al., 2009b) or questionable practices concerning equity issues (Booher-Jennings, 2005). By contrast, DBDM has the potential to foster the attainment of equity goals (Dodman et al., 2019) and, among other benefits, mitigate absenteeism (Balfanz & Byrnes, 2018). The important preconditions of appropriate DBDM include data literacy and the motivational beliefs of teachers regarding data use (Datnow & Hubbard, 2016; Filderman et al., 2022; Prenger & Schildkamp, 2018). Thus, it has been argued that these factors should be addressed at all stages of the teaching career, beginning at the pre-service level by building awareness of and skills in using data and related tools (Beck & Nunnaley, 2021; Mandinach & Gummer, 2013).

Against this background, the objective of the present study is to investigate the effects of a data literacy intervention with pre-service teachers on their data literacy test results and motivational beliefs about data use. Initially, we offer an overview of approaches to foster DBDM among teachers by providing conceptual insights into data literacy and motivational beliefs about data use, along with corresponding research results. Based on that, we formulate research questions and describe the present study.

7.1. Fostering data-based decision making

DBDM has been defined as “systematically analysing existing data sources within the school, applying outcomes of analyses to innovate teaching, curricula, and school performance, and, implementing (e.g., genuine improvement actions) and evaluating these innovations” (Schildkamp & Kuiper, 2010, p. 482). In recent decades, research on DBDM has sought to identify factors fostering or hindering successful data use in education and to collate them on different levels: organizational-contextual, data or user (Hoogland et al., 2016; Schildkamp et al., 2014, 2017). Research focusing on teachers as data users has consistently reported that their skills, knowledge and competencies regarding data use – which can be summarized as data literacy – are the key preconditions (Hoogland et al., 2016; Schildkamp et al., 2014, 2017). In addition to their data literacy, teachers’ motivational beliefs about data use have been highlighted as important (Datnow & Hubbard, 2016; Hoogland et al., 2016; Prenger & Schildkamp, 2018). Hence, a short overview of conceptual models and empirical findings concerning data literacy and motivational beliefs about data use among teachers is provided in the next sections.

7.1.1. Teachers' data literacy

Mandinach and Gummer (2016) proposed a detailed conceptual framework on data literacy for teachers (DLFT), which is conceptualized as follows:

> "the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data (assessment, school climate, behavioral, snapshot, longitudinal, moment-to-moment, etc.) to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn." (Mandinach & Gummer, 2016, p. 367)

1) Identify problems and frame questions	First, a problem concerning, for example, the curriculum or an aspect of instruction, has to be identified and concrete questions formulated in consideration of the specific school context.
2) Use data	This is the largest component and covers actual data use. Appropriate data (sources) have to be identified or generated and data has to be analysed and understood, such as by aggregating and disaggregating data.
3) Transform data into information	Data has to be interpreted in consideration of the specific context, which can include making inferences, generating explanations and drawing conclusions.
4) Transform information into a decision	Decisions for the next (instructional) steps have to be planned and conducted based on the information generated.
5) Evaluate outcomes	As a final component, the impact of the decisions and adjustments regarding the original question or problem has to be evaluated, such as by considering both desired and unintended effects.

Tabelle 3. The five components of data literacy for teachers (Mandinach & Gummer, 2016)

The DLFT framework consists of five components (see Tabelle 3) oriented towards different parts of the cyclical and sequential use of data. However, Mandinach and Gummer (2016) highlight that the components also have some degree of interrelation and do not necessarily have to be followed sequentially; for example, one might return to an earlier component before evaluating outcomes. In addition, there are dispositions, habits of mind and other factors (notably, the belief that data can be helpful or that all students can learn) that are closely

connected to data literacy, although they are not directly part of data literacy in their framework, because they refer to teaching in general. The same authors point out that the use of any kind of data – not just assessment data – must be placed within a content- and classroom-specific context. Therefore, data literacy must be integrated with general teacher knowledge like content knowledge, pedagogical content knowledge and general pedagogical knowledge.

Although teachers' data literacy can be considered a key factor in successful data use, many studies on the topic have revealed low levels of competencies and skills in practice (Gelderblom et al., 2016; Huguet et al., 2014; Kippers et al., 2018; J. A. Marsh, 2012; Means et al., 2011; van der Kleij & Eggen, 2013; Zeuch et al., 2017). Teachers consistently report struggling with how to use data to adequately inform instructional decisions (Espin et al., 2018; Hoogland et al., 2016; Reeves & Chiang, 2018; van Geel et al., 2017). This raises the issue of how both in- and pre-service teachers' data literacy can be fostered (see section 7.1.3).

7.1.2. Teachers' motivational beliefs about data-based decision making

Broadly speaking, from a motivational theory point of view, it is not only people's knowledge and skills that are important but also their motivational beliefs, values and goals (Eccles & Wigfield, 2002). Hence, researchers have investigated the motivational beliefs, values and goals of teachers regarding DBDM in a general sense. Several systematic reviews and meta-analyses highlight the importance of data users' beliefs as a key precondition of successful DBDM (Coburn & Turner, 2011; Datnow & Hubbard, 2016; Filderman et al., 2022; Hoogland et al., 2016). However, one can find heterogeneous conceptualizations and approaches concerning motivational beliefs about DBDM in the field. To give a structured overview of those studies and clarify their focuses, we follow the approach of four broad perspectives of motivational theories, that Eccles and Wigfield (2002) worked out in their review: the first focuses on "beliefs about competence and expectancy for success", the second on "reasons why individuals engage in different activities", the third on "theories that integrate expectancy and value constructs" and the fourth on "theories integrating motivation and cognition" (pp. 109–110).

Regarding the first category, teachers' self-efficacy concerning DBDM has been investigated in many studies (Albritton & Truscott, 2014; Dunn et al., 2013; Hamilton & Reeves, 2021; Lembke et al., 2018; Oslund et al., 2021; Pierce & Chick, 2011; Prenger & Schildkamp, 2018; Reeves et al., 2020; van der Scheer & Visscher, 2016; Walker et al., 2018). In addition, research focusing on teachers' self-assessed data literacy can be assigned to this category (Reeves & Chiang, 2018, 2019; Reeves & Honig, 2015; Supovitz & Sirinides, 2018). The second category covers intrinsic motivation, interest and goal theories and research, but we

could not find relevant research on DBDM. The third category encompasses expectancy value theories that combine expectancy (beliefs in one's ability concerning a domain or task), values regarding a task (attainment, intrinsic and utility value) and cost (Eccles et al., 1983). In DBDM research, there are studies that fall into this category but only assess specific aspects of the value component by focusing on, for example, the importance of test data (Supovitz & Sirinides, 2018) or beliefs concerning responses to intervention (Castillo et al., 2016). By contrast, Reeves and Honig (2015) and Reeves and Chiang (2017, 2018, 2019) investigated both self-efficacy about DBDM and different beliefs about assessment, but they did not address beliefs about data or DBDM in general. The only study to focus on both the expectancy and value components of DBDM in teaching is Thoren et al. (2020), even though it appears important to investigate those two aspects simultaneously, not only from an additive perspective but also considering the interaction of expectancy and value (Nagengast et al., 2011). No studies were found concerning the fourth category of theories and research integrating motivation and cognition regarding DBDM. Thus, most research on investigating teachers' beliefs about their abilities concerning DBDM (the first category) or aspects of expectancy and value (the third). Due to the importance of motivational beliefs for successful DBDM, a remarkable amount of research addresses the issue of fostering motivational beliefs about DBDM.

7.1.3. Fostering teachers' data literacy and motivational beliefs

Taking up the issue of fostering teachers' data literacy and motivational beliefs, Filderman et al. (2022) carried out a meta-analysis that evaluated trainings on data literacy and beliefs about DBDM. Overall, they found strong positive effects on knowledge and skills (g = 0.67; 95% CI = [0.40, 0.93]) and moderate effects on beliefs (g = 0.48; 95% CI = [0.17, 0.79]). Untangling the latter effects into self-efficacy and value components, they report ranges of effect sizes for self-efficacy from g = -0.74 to g = 3.68 and for value from g = -0.43 to g = 1.25. Filderman et al. (2022) interpret these results to conclude that teachers' beliefs on "the value of data may be less malleable than their beliefs in their own ability to use data" (p. 11).

However, only seven of the 33 studies in Filderman et al.'s (2022) corpus include pre-service teachers, so inferences regarding fostering their data literacy and beliefs should be drawn carefully, and more research beyond their meta-analysis is merited. Focusing on pre-service teachers' data literacy, Reeves and Honig (2015) conducted a data literacy intervention study on the use of summative assessments and found moderate positive effects on pre-service teachers' data literacy skills (d = 0.60, p < .001), while Reeves and Chiang (2017) found a small positive effect in their study (d = 0.34, p < .05). Merk et al. (2020) developed a data literacy test and data literacy intervention for pre-service teachers focusing on data use in a

narrow sense (data collection, transformation, reduction and interpretation). They reported a large positive and significant effect of the intervention. Studies focusing on expectancy and value beliefs have reported mixed effects: by and large, intervention studies on self-efficacy regarding DBDM indicate moderate to large positive effects for self-efficacy, with the limitation that not all effects are statistically significant (Reeves & Chiang, 2017, 2018, 2019; Reeves & Honig, 2015). Concerning self-assessed data literacy, Thoren et al. (2020) found small to large significant positive effects. Overall, studies on the value of assessment data report moderate but generally not significant positive effects (Reeves & Chiang, 2017, 2018, 2019; Reeves & Honig, 2015) and slight and non-significant changes in the values of DBDM in general (Thoren et al., 2020). Taken together, the results in the literature, especially with pre-service teachers on motivational beliefs about DBDM, remain largely inconclusive due to many non-significant studies that do not differ between evidence of absence of an effect (null hypothesis) and the absence of evidence (Dienes, 2016). To sum up, there is some evidence that data literacy skills can be fostered effectively in (pre-service) teachers, but it is largely unclear whether this holds true for motivational beliefs, especially the value component.

7.1.4. Research questions

As outlined above, data literacy and motivational beliefs about DBDM are crucial prerequisites for successful DBDM and should thus be addressed in pre-service teacher education (Beck & Nunnaley, 2021). Although intervention studies concerning data literacy and motivational beliefs about DBDM among pre-service teachers offer some valuable approaches, further investigation into how to effectively foster data literacy in brief trainings is needed, especially regarding rigorous research designs and interventions (Datnow & Hubbard, 2016; Hoogland et al., 2016; Reeves & Chiang, 2018). Regarding motivational beliefs, previous research has generally focused on either expectancy or value with an emphasis on assessment (data) rather than on DBDM in general.

The present study helps fill these gaps by exploring the effects of a short and easily implementable intervention with pre-service teachers on data literacy and motivational beliefs about DBDM. To investigate motivational beliefs about DBDM in their entirety, they were conceptualized as constructs of expectancy and value (Eccles & Wigfield, 2002). First, we analyse the within- and between-person structure of motivational beliefs about DBDM (RQ1), as a prerequisite for further analyses. Second, we investigate the effects of the intervention on the improvement of data literacy and motivational beliefs (RQ2). Based on previous studies and theoretical concepts of the development of data literacy, we assume a positive effect of the intervention on data literacy (Beck & Nunnaley, 2021, Filderman et al., 2022; Merk et al., 2020; Reeves & Honig, 2015). Concerning the changeability of components of motivational

beliefs in the context of data use, the state of research is mixed and partly inconclusive (Filderman et al., 2022; Reeves & Chiang, 2018; Reeves & Honig, 2015; Thoren et al., 2020). Therefore, we formulate informative hypotheses and use Bayesian statistics (see sections 7.2.4 and 7.3.2.2) to provide evidence not only against but also in favour of null effects (Dienes, 2016).

7.2. Method

7.2.1. Sample

The study was designed as a randomized controlled trial with a wait list control group (CG). Participants were recruited from five educational courses for pre-service secondary teachers at a large university in Germany. We conducted two studies: in the pilot (Study 1), the structure of motivational beliefs about DBDM was investigated (N = 34; $M_{semester}$ = 5.7, $SD_{semester}$ = 1.3; 67% female; 64% no STEM subject). In Study 2, the main study, 136 course participants agreed to take part in the project, and 132 were present for all measurement occasions. All students were randomly drawn into one of two conditions ($N_{treatment}$ = 68; $N_{control}$ = 64). The treatment group (TG) received the intervention first, while the CG received the intervention afterwards. During the intervention, the CG took part in the standard curriculum of the seminar and received content about differentiated instruction, its advantages and challenges and how to implement differentiated instruction. Furthermore, students learned about the importance of formative assessment and feedback in this context. The participants were mostly at the end of their Bachelor of Education studies ($M_{semester}$ = 5.5, $SD_{semester}$ = 1.4), predominantly female (63%) and not studying a STEM subject (61%).

7.2.2. Intervention

7.2.2.1. Content

We base our intervention on the DLFT framework by Mandinach and Gummer (2016) and the corresponding aspects of data literacy training in Filderman et al. (2022). All elements of DBDM suitable for inclusion in pre-service teacher education were taken into account (Beck & Nunnaley, 2021; Merk et al., 2020). The intervention focused on building the basic concept of data literacy and considers different types of data, such as assessment and self-evaluation data (Mandinach & Schildkamp, 2021a). The main emphasis was on the data analysis phases, like the “use data”, “transform data into information” and “transform data into a decision” components in the DLFT framework in Mandinach and Gummer (2016).

First, the fundamental ideas of DBDM were introduced, along with the basics on data collection, measurement of variables and criteria of measurement quality. In this part, opportunities and challenges concerning DBDM were introduced and discussed, such as

using student feedback as an opportunity to capture students' perceptions of instructional quality and use them for improvement. Second, data transformation and data reduction techniques were outlined, using quantitative characteristics, qualitative terms and graphs (e.g., the distribution of grades within different courses). Third, methods of data transformation (e.g., scaling, centring and standardising) and additional methods of data reduction (comparison of means and correlation of variables) were described. Details of the intervention content and its connections to Mandinach and Gummer (2016) and Filderman et al. (2022) appear in Appendix Table A1, and examples of the activities and materials can be found on the Open Science Framework (OSF; https://osf.io/9c6h7/).

7.2.2.2. Design principles and activities

The intervention took place in the middle of the semester with approximately 6 h of learning time. It was part of a larger course designed in the style of a massive open online course to take place asynchronously and thus enable self-regulated learning by the students. The intervention consisted largely of video-based instruction and supporting materials like interactive exercises programmed with the R package Shiny and worked-out examples, such as a case study of student feedback about instructional quality. The students used synthetic data, based mainly on the examples provided. Details of the learning activities are provided in Appendix Table A1.

The design of the intervention drew on the findings of a metaanalysis about high-quality features of teacher data literacy training (Filderman et al., 2022) and other work about fostering data literacy among pre-service teachers through online interventions (Reeves & Chiang, 2018; Tallent-Runnels et al., 2006). Wherever feasible in the online context, high-quality training features were implemented. To provide a highly structured and coherent course, recommendations about the subject matter for each week and time limits for each topic were provided. To mimic the presence of a facilitator or coach, students were able to ask questions over email and in a forum at any time. Learner-instructor interaction was also realized through regular automated individual feedback that relied on the results of quizzes for each part. Furthermore, all students received individualised feedback at the end of the intervention about their learning gains in data literacy and changes in beliefs. That feedback was intended to serve as a basis for the students to complete a written reflection on the learning process at the end of the intervention. Active learning and collective participation through learner-learner interaction was enabled by peer feedback and the forum.

7.2.2.3. Treatment check

The treatment check was investigated using the three quizzes for each part of the intervention, on which the students averaged 83%–90% correct answers. We could thus assume adequate implementation fidelity.

7.2.3. Instruments

7.2.3.1. Motivational beliefs

We used a slightly adapted instrument from Thoren et al. (2020) to assess motivational beliefs about DBDM, as it is based on a widely accepted theoretical framework (Eccles & Wigfield, 2002) and covers a broad range of Mandinach and Gummer's aspects of data literacy (see section 7.1.1). This instrument assesses the classic components of Eccles and Wigfield's (2002) expectancy value theory: beliefs in competence and self-efficacy cover the expectancy part, whereas task values refer "to the reasons for engaging in a specific task" (Plante et al., 2013, p. 67). These task values are divided into intrinsic value (enjoyment while carrying out the task), attainment value (personal importance of high task performance), utility value (relation to future goals) and cost (associated psychological expense). However, respondents were not asked to rate their expectancy of and values regarding DBDM as a whole but for 13 specific parts (e.g., formulating a research question, generating data, deriving instructional consequences). Thoren et al.'s (2020) instrument uses item stem and answer anchors to operationalize the motivational beliefs and the items themselves to operationalize each component of data literacy. For example, the stem "please indicate how much you enjoy the following" introduced items such as "formulating a question which is relevant for practice and can be answered with data" and "collecting appropriate data by using instruments". The answer anchors ranged from 1 (is something that I do not enjoy) through six (is something that I really enjoy). The full instrument, translated from German, can be found on the OSF (https://osf.io/9c6h7/).

7.2.3.2. Data literacy test

We used a short data literacy test from Merk et al. (2020) that focuses on aspects of data literacy that are especially relevant for pre-service teachers and covers in particular two components of Mandinach and Gummer's DLFT (2016): data use (understanding data properties, manipulating data, aggregating data) and the transformation of data into information (understanding and using data displays and representations, using statistics). The test contains 10 items and can be found on the OSF (https://osf.io/9c6h7/). The exact assignment of the test items to the components can be found in Table 2 in Merk et al. (2020).

7.2.4. Statistical analysis

Our research questions focus on the structure of motivational beliefs about DBDM both within and between individuals (RQ1) and any changes in those beliefs induced by an intervention (RQ2). To investigate RQ1, one has to consider the complexity of the instrument we used: these beliefs are a) regarded as made up of five components (expectancy, enjoyment, attainment value, utility, cost), b) viewed as related to 13 aspects of DBDM and c) measured at two timepoints. Hence, every item can be precisely mapped to one component of motivational beliefs and one component of DLFT, resulting in a cross-classified or imperfect hierarchical data structure (see Abbildung 15; Snijders & Bosker, 2012). As it is viewed as almost impossible (H. W. Marsh, 2007) to achieve an acceptable fit for instruments of such complexity (five or more factors, each with more than five items and more than 50 items overall), we began by treating the five value components as nested within persons and used cluster robust exploratory factor analysis (CREFA) on the pre-treatment data. This also increases statistical power because the number of individuals equals the sample size at level 2.

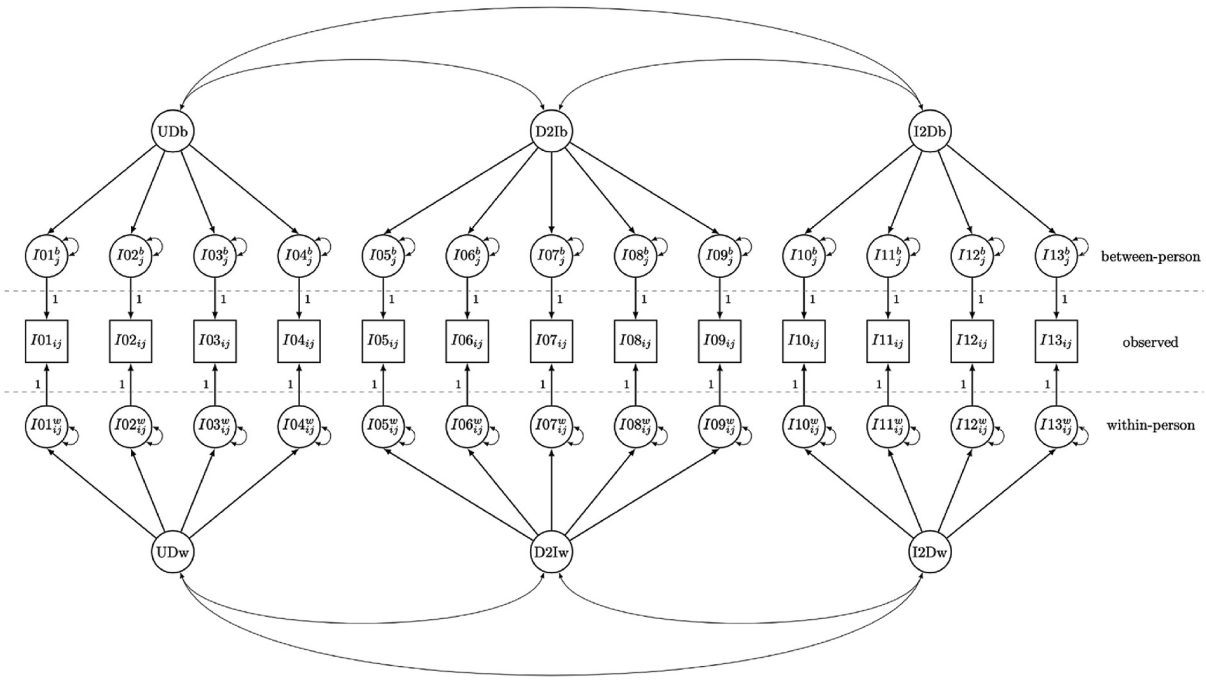


Abbildung 15. Multilevel structure of the dimensions of motivational beliefs about DBDM

Notes. UD = “Identify problems and frame questions and use data”; D2I = “Transform data into information”; I2D = “Transform information into a decision and evaluate outcomes”. For clarity, only one component is depicted.

To confirm the factorial structure found through this approach and to investigate the hierarchical variance decomposition, we applied cluster robust confirmatory factor analysis (CRCFA) and multilevel confirmatory factor analysis (MLCFA; Mehta & Neale, 2005) to the pre-treatment data from Study 2; see section 7.3.2.1 for details. We did not conduct a priori

power analyses for the MLCFA, as how to accurately choose which among so many parameters (e.g., item intercepts, slopes) to specify a priori was not clear at all.

To model the effects of the intervention, we specified multi-group trivariate latent change score models (LCSMs; Kievit et al., 2018; McArdle & Grimm, 2010) to go beyond effect sizes for mean changes and to grasp information about variance in changes, their relations to initial scores and the latent correlations of change scores. To provide evidence not only against but also in favour of null effects (Dienes, 2016), we applied a Bayesian hypothesis testing framework for structural equation modelling (Gu et al., 2019), as previous research results concerning the effect of interventions are mixed (see section 7.1.3.). A power approximation for these models using Monte Carlo approaches implemented in the WebPower R package (Zhang & Yuan, 2018) resulted in sufficient (frequentist) power (>0.85) for a sample size of 100 and moderate effect sizes. All analyses are documented on the OSF (https://osf.io/9c6h7/).

7.3. Results

7.3.1. Study 1 (pilot study)

7.3.1.1. Factor structure

To investigate the structure of motivational beliefs about DBDM, we first applied CREFA with geomin rotation using the Mplus framework (Muthén & Muthén, 2017). A solution with three factors outperformed models with one, two and four factors (see Tabelle 4).

Number of factors	χ^2 (df)	*p*	CFI	TLI	RMSEA
1	308.686 (65)	< .001	.74	.69	.14
2	194.479 (53)	< .001	.85	.78	.12
3	81.304 (42)	< .001	.96	.92	.07
4	81.942 (32)	< .001	.95	.87	.09

Tabelle 4. Model fit comparison cluster robust exploratory factor analysis

The factor loadings of the final model (see Appendix Table A2) were in line with Mandinach and Gummer's DLFT (2016): identify problems, frame questions and use data (factor one with four items), transform data into information (factor two with five items) and transform information into a decision and evaluate outcomes (factor three with four items). The only exception was item two ("Understand the purpose of different data sources [e.g., comparative tests, self-evaluation data]"), which showed substantial loadings on two factors. To ensure substantial coherence with the DLFT construct, we assigned item two to factor one. All three

factors showed good reliability ($0.77 \leq \omega \leq 0.92$) for each of the five components of motivational beliefs (see Appendix Table A3).

7.3.2. Study 2 (main study)

7.3.2.1. Factor structure and reliability

To confirm the factorial structure found in Study 1, CRCFA with participants modelled as clusters and MCFA were applied to estimate congeneric measurement models. Both CRCFA and MCFA showed sufficient model fit (see Tabelle 5) after freely estimating two residual covariances (selected by modification indices), except for $SRMR_{between}$ in the MCFA model. This insufficient fit at the between-subject level can be interpreted as evidence for the hypothesis that teachers do not generalize the distinction between the three factors over the five components of motivational beliefs.

Model	χ^2 (df)	*p*	CFI	TLI	RMSEA	$SRMR_{within}$	$SRMR_{between}$
CRCFA	287.72 (60)	< .001	.95	.94	.08	.05	
MCFA	624.74 (126)	< .001	.95	.94	.06	.05	.26

Tabelle 5. Model fit cluster robust and multilevel confirmatory factor analysis

Hierarchical variance decomposition was investigated by calculating intraclass correlations (ICCs), which are greater than zero ($0.00 \leq ICC \leq 0.11$) for most variables. The low level of variance at the between-person level indicates that participants distinguish strongly between DLFT components (factors); therefore, a specific modelling is necessary. All scales of motivational beliefs about DBDM show good reliability ($0.80 \leq \omega \leq 0.93$) for each of the three factors in all belief components (see Appendix Table A4). The reliability analysis of the data literacy test using ordinal Cronbach's α (Gadermann et al., 2012) resulted in sufficient reliability for both pretest ($\alpha = 0.78$) and posttest ($\alpha = 0.72$).

7.3.2.2. Effects of the intervention

First, we visualized the descriptive data to obtain a first impression (Abbildung 16). The results of Bayesian hypothesis testing and latent change scores are described below. Abbildung 16 shows a positive effect of the intervention on the data literacy test in the TG and no relevant effect in the CG. Overall, most motivational belief components were rated positively on average and showed negligible change between the TG and CG, except for changes in self-efficacy and certain factors in the components of motivational beliefs. The intervention showed

increased self-efficacy for all dimensions; that was also present in the CG, albeit to a lesser extent. Other components of beliefs impacted by the intervention were attainment and utility in factor one (using data) and slightly increased enjoyment and decreased perceived costs in factor two (transform data into information).

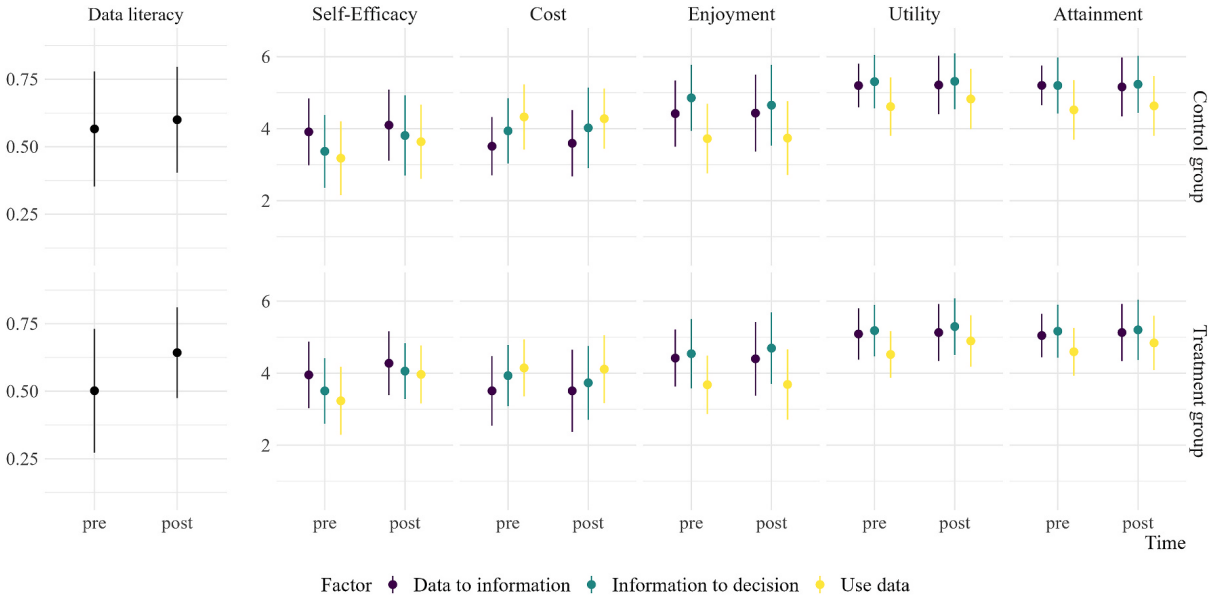


Abbildung 16. Effects of the intervention on data literacy and motivational beliefs (*M* ± 1*SD*)

Second, multi-group LCSMs were applied to estimate the effects of the intervention (groups) and changes in the three factors of each component of motivational beliefs. Therefore, a two-group trivariate LCSM was separately estimated for each component of motivational beliefs (see Abbildung 17).

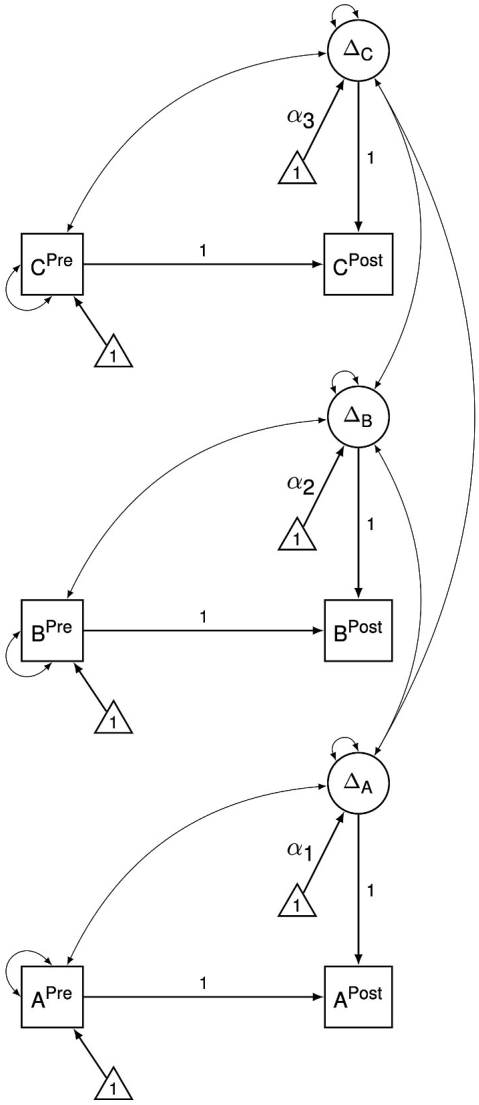


Abbildung 17. Multi-group trivariate latent change score model

Notes. $A^{Pre}/B^{Pre}/C^{Pre}$ and $A^{Post}/B^{Post/}C^{Post}$ are the pre- and post-measurements, respectively, of factor one ("Identify problems and frame questions and use data" [A]), factor two ("Transform data into information" [B]) and factor three ("Transform information into a decision and evaluate outcomes" [C]). Meanwhile $\Delta_A/\Delta_B/\Delta_C$ are the latent difference scores between pre- and posttests for each factor, and $\alpha_1/\alpha_2/\alpha_3$ are the intercepts respectively means of the latent difference score variables. For clarity, only one group and one component are depicted.

The standardised latent change scores and results of Bayesian hypothesis testing are presented in Abbildung 18. For each component, three informative hypotheses regarding differences between the TG and CG were specified. Hypothesis 1 (H1: $\alpha_{CG} = 0$ & $\alpha_{TG} = 0$), the null hypothesis, assumes all mean latent change scores to be zero and no differences between the two groups. Hypothesis 2 (H2: $\alpha_{CG} = 0$ & $\alpha_{CG} < \alpha_{TG}$), the intervention effect hypothesis, assumes mean latent change scores in the CG to be zero and smaller than the TG mean change score and was formulated to detect a potential positive effect of the intervention. Hypothesis 3 (H3: $\alpha_{CG} < \alpha_{TG}$) assumes the mean latent change scores in the CG to be smaller than in the TG. It was formulated to detect potential changes in the CG, such as those due to transfer effects caused by the wait list CG design (see section 7.2.1). The

hypotheses were evaluated using approximate adjusted fractional Bayes factors, which are equivalent to posterior model probabilities (PMPs). The latter can range between zero and one; combined, they add up to one for all hypotheses under examination. For better understanding, they are visualized in pie charts (see Appendix Table A5 for exact values). PMPs can be interpreted as measures of support in the data for each hypothesis. The intervention has a substantial effect on the data literacy test results in the TG but no effect in the CG, which is supported by a high PMP (.82) for H2. According to the motivational beliefs' components, the LCSMs generally show no substantial change in either CG or TG for cost and enjoyment, and the corresponding PMPs provide strong evidence for H1. By contrast, pre-service teachers in both the TG and CG reported increased self-efficacy, while the mean latent change scores are higher in the TG than in the CG. This is supported by a PMP near one for H3. In the TG, all factors show substantial change ($0.31 \leq \alpha \leq 0.72$), whereas in the CG this is only true for factors one ($\alpha_1 = 0.42$) and three ($\alpha_3 = 0.43$). Fairly inconclusive results are found for utility and attainment. Changes in the utility to use data and attainment are only relevant for factor one in the TG (attainment: $\alpha_1 = 0.31$; utility: $\alpha_1 = 0.47$), and the values of the PMP cannot be interpreted as offering strong support for any of the hypotheses. Thus, there is no completely consistent effect of the intervention on beliefs. Nevertheless, the intervention does show a positive impact on self-efficacy and factor one (use data) of utility. The results show that the three factors enable differentiated analyses of motivational beliefs, depending on the different steps in the data use process.

Mean latent change score	Group	Data literacy	Self-efficacy	Cost	Enjoyment	Utility	Attainment
α_1	TG	0.76	0.72	-0.03	0.01	0.47	0.31
	CG	0.14	0.42	-0.02	-0.01	0.23	0.13
α_2	TG		0.60	-0.02	0.19	0.06	0.11
	CG		0.19	0.11	-0.23	0.02	-0.06
α_3	TG		0.31	0.00	-0.02	0.16	0.04
	CG		0.43	0.13	-0.02	0.01	0.03
Posterior Model Probabilities							

Notes. TG = treatment group, CG = control group; α_1 = mean latent change score of factor one "Identify problems and frame questions & use data"; α_2 = mean latent change score of factor two "Transform data into information"; α_3 = mean latent change score of factor three "Transform information into a decision & evaluate outcomes". All mean latent change scores are standardized and can be interpreted as effect sizes. Posterior Model Probabilities concerning hypotheses with respective colors.

H1: $\alpha_{1\,CG} = \alpha_{2\,CG} = \alpha_{3\,CG} = 0$ & $\alpha_{1\,TG} = \alpha_{2\,TG} = \alpha_{3\,TG} = 0$

H2: $\alpha_{1\,CG} = \alpha_{2\,CG} = \alpha_{3\,CG} = 0$ & $\alpha_{1\,CG} < \alpha_{1\,TG}$ & $\alpha_{2\,CG} < \alpha_{2\,TG}$ & $\alpha_{3\,CG} < \alpha_{3\,TG}$

H3: $\alpha_{1\,CG} < \alpha_{1\,TG}$ & $\alpha_{2\,CG} < \alpha_{2\,TG}$ & $\alpha_{3\,CG} < \alpha_{3\,TG}$

Abbildung 18. Results of multi-group trivariate latent change score models of motivational beliefs and the univariate latent change score model of the data literacy test

Interrelations between the changes measured by the correlations of the latent change scores reveal medium to large latent correlations (Appendix Table A6). The greater a correlation, the

more similar the change between the factors within a component of motivational beliefs. This means that although there are different change processes within components of motivational beliefs, those changes tend to be similar.

7.4. Discussion

7.4.1. Summary of findings

Data has become commonplace in schools and the educational system as a whole, and DBDM has the potential to inform pedagogical decisions and improve educational practices. Therefore, some scholars have argued that data literacy and beliefs about DBDM should be part of pre-service teacher education (Beck & Nunnaley, 2021; Mandinach & Gummer, 2016). The aim of the studies presented above was to explore the effect of a short and easily implementable intervention to foster specific components of pre-service teachers' data literacy and to investigate the effect on beliefs about DBDM.

RQ1 focused on the structure of motivational beliefs about DBDM that are based on the expectancy value theory (Eccles & Wigfield, 2002) both within and between individuals. Study 1 revealed a new three-factor structure of Thoren et al.'s (2020) motivational beliefs about DBDM, which are in line with DLFT (factor one, use data; factor two, transform data into information; factor three, transform information into a decision and evaluate outcomes; Mandinach & Gummer, 2016). The structure was confirmed in Study 2 and thus enabled separate investigations of the different parts of DLFT.

Study 2 confirmed the hypotheses behind RQ2 as follows. The intervention has a large effect on aspects of data literacy and a large effect on self-efficacy. Concerning data literacy, the results are in line with Merk et al. (2020); the online intervention has a slightly lower effect on the data literacy test than a similar intervention with face-to-face teaching. Positive effects on self-efficacy have also been found in other studies with pre-service teachers (Reeves & Chiang, 2018; Reeves & Honig, 2015; Thoren et al., 2020), but they were not found in all intervention studies (Filderman et al., 2022). The positive effects on self-efficacy in the CG are presumably a transfer effect due to the seminar content about formative assessment in which the CG took part while the TG took part in the intervention.

The present study shows that, on average, pre-service teachers have positive motivational beliefs regarding DBDM. On the basis of Bayesian hypothesis testing, we found strong evidence that the value components enjoyment and cost are largely unchanged by the intervention. The results on beliefs about DBDM are comparable to those in Thoren et al. (2020), except that those authors report a stronger decrease in cost. Similar results are found

in Reeves and Honig (2015), who report small changes in two of 11 motivational belief scales, which were mostly related to assessment data. Contrary to our findings, Reeves and Chiang (2018) report changes in pre-service teachers' anxiety and assessment beliefs. However, they focused on assessment beliefs, so the results are not directly comparable to generic beliefs about DBDM and various data sources in our study.

7.4.2. Limitations

The intervention in the present study is based on important design elements of successful online interventions, like a highly structured course, elements of online learner-instructor interaction (e.g., regular automated individual feedback) and learner-learner interaction (e.g., peer feedback; Filderman et al., 2022; Reeves & Chiang, 2018). But the intervention does not fulfil all high-quality training features of data use training (Ansyari et al., 2020; Filderman et al., 2022), such as long duration and coaching, which could not be effectively imitated in a brief online intervention. The unique features were the short duration of the intervention and the design oriented towards massive online open courses without face-to-face interaction with instructors. Our intervention was not explicitly designed to impact beliefs, which could be seen as a limitation. Another limitation is the lack of a follow-up test, so the persistence of the induced effects remains unclear. Since the study focuses on pre-service teachers, it is unknown whether the data literacy skills and mostly positive beliefs about DBDM will survive years of daily work in schools. Although there are strong indicators for a sufficient treatment check, this conclusion is based only on quizzes, with no other indicators available. Another limitation is the partial curricular validity of the data literacy test, as some content (identifying problems and questions; transforming information into decisions and evaluation of outcomes) was only marginally included in the intervention, and items to measure the transformation of information into decisions and evaluation of outcomes were not included.

7.4.3. Implications and future research

Despite these few limitations, the study provides strong evidence by using a randomized controlled trial to go beyond self-reports on data literacy and relies on a strong theoretical foundation of motivational beliefs. The short intervention has the potential to foster aspects of data literacy and self-efficacy about DBDM but not value components of beliefs about DBDM in general. Since data literacy is not yet part of most teacher education programs, a short intervention of the kind used here seems relatively easy to implement. With regard to in-service teachers, Filderman et al. (2022) report similar effects in their meta-analyses. Furthermore, our results are in line with the conclusion those authors draw that beliefs about the value of DBDM appear to be less malleable than self-efficacy beliefs about DBDM. One explanation in terms of our findings is that our participants already had mostly positive beliefs.

Whether beliefs about DBDM can be influenced by interventions – and if so how – remain largely open questions, but it seems possible, at least for certain parts of beliefs. Filderman et al. (2022) suggest explicit training on the rationale for DBDM, especially to foster long-term improvements in practice (see also Hoogland et al., 2016). This should be addressed in further intervention studies, especially with regard to the (development of the) interaction between expectancy and value (Nagengast et al., 2011). Furthermore, the relationship between the development of self-efficacy and test performance should be investigated in greater depth. An open question is the role of teacher educators and curricula. Adequate data use might also be a challenge for teacher educators, and the values of DBDM and buy-in to data use might be not a given among that important group. Therefore, the data literacy and beliefs of teacher educators at different stages of the development of teachers' data literacy (Beck & Nunnaley, 2021; Mandinach & Gummer, 2016) and their effects on (pre-service) teachers' beliefs and data literacy education should be investigated.

Further research could also focus on whether data literacy should be taught generically or using data-specific approaches that involve, for example, assessment data or self-evaluation data. Generic data literacy of the kind used in our intervention has the advantage of being connectable to a wide range of data sources. However, the authenticity and context of data are also important (Beck & Nunnaley, 2021). This leads to a further open question: how can action planning, implementation and evaluation of outcomes be integrated into pre-service teacher education, if no own class and teaching experience is available? Despite these questions, the present study provides strong evidence of the effects of a short data literacy intervention on pre-service teachers. Against the background of increasing availability of data and repeated calls to use data for teaching, it offers a promising approach to integrate data literacy into teacher education.

8. Data-based decision making in einer digitalen Welt: Data literacy von Lehrpersonen als notwendige Voraussetzung (Artikel 4)

Bez, S., Tomasik, M. J., & Merk, S. (2023). Data-based decision making in einer digitalen Welt: Data Literacy von Lehrpersonen als notwendige Voraussetzung. In K. Scheiter & I. Gogolin (Hrsg.), *Bildung für eine digitale Zukunft* (S. 339–362). Edition ZfE. Springer VS. https://doi.org/10.1007/978-3-658-37895-0_14

Stichworte
Datenbasierte Schul- und Unterrichtsentwicklung, Data literacy, Digitalisierung, Lehrpersonen

Abstract
Im Zuge der Digitalisierung entstehen für die datenbasierte Schul- und Unterrichtsentwicklung nicht nur neue Daten(quellen), entsprechende Technologien machen darüber hinaus auch die Transformation, Aggregation, Verknüpfung und Dissemination vergleichsweise einfach möglich. Allerdings garantiert das weder die gewinnbringende Nutzung dieser Daten, noch resultiert es ohne weiteres in verbesserten Lernprozessen und -ergebnissen, etwa bezüglich einer verstärkten individuellen Förderung. Dies führt zur These dieses konzeptuellen Beitrags: Um die Potenziale für die datenbasierte Unterrichtsgestaltung und -entwicklung, die mit der Digitalisierung im Bildungswesen einhergehen, ausschöpfen und potenzielle dysfunktionale Wirkungen minimieren zu können, ist die *data literacy* der Lehrkräfte eine notwendige Voraussetzung. Nach einer Begriffs- und Konstruktklärung werden im Beitrag anhand des Forschungsstandes zur data literacy von Lehrpersonen die notwendigen Voraussetzungen für das Gelingen skizziert. Daraufhin wird analysiert, welche Potenziale aktuelle digitale Innovationen für eine verstärkte Realisierung individueller Förderung bieten, und beispielhaft veranschaulicht, wie zentral die data literacy von Lehrpersonen für die Hebung dieses Potenzials ist. Abschließend werden Forschungsperspektiven sowie Implikationen für die Praxis formuliert.

8.1. Einleitung

Wie in nahezu allen gesellschaftlichen Bereichen hält auch die Digitalisierung immer mehr Einzug in Schule und Unterricht: Dies betrifft bspw. den (fach)didaktischen Einsatz digitaler Medien im Unterricht, die Implementation organisationsbezogener Tools wie Lernmanagementsysteme und digitale Klassenbücher sowie die Nutzung von Innovationen wie technologiebasierte formative Assessmentsysteme usw., wenn auch in unterschiedlicher Intensität auf den Ebenen der Lehrpersonen und Einzelschulen (Hartong et al., 2019; Jude et al., 2020). Auch wenn viele der aufgeführten Beispiele zunächst als eine reine Ersetzung bisheriger analoger Prozesse erscheinen, wie etwa die Erfassung von Fehlzeiten in digitaler Form anstelle von Stift und Heft, entstehen so bereits vielfältigste neue Daten(quellen). Allerdings macht die Digitalisierung unter der Voraussetzung der Maschinenlesbarkeit der Daten zunächst die Verfügbarkeit, Transformation und Teilbarkeit überhaupt möglich bzw.

sehr einfach (wenn etwa nicht mehr ein einzelnes Papierexemplar des Klassenbuchs von allen pädagogischen Fachkräften einer Klasse gemeinsam genutzt werden muss, sondern alle in Echtzeit auf dieselben Daten Zugriff haben). Zudem erlaubt das digitale Format die Transformation, Aggregation und Verknüpfung der Daten, die analog nicht oder nur höchst aufwändig möglich wären und sehr wahrscheinlich mit Qualitäts- und Validitätsverlusten einhergehen würden (z.B. lassen sich bestimmte Muster von Fehlzeiten bzw. Schulabsentismus auf Individual-, Klassen- oder Schulebene finden). Für die datenbasierte Schul- und Unterrichtsentwicklung ergeben sich dadurch vielversprechende Möglichkeiten. Denn der Nutzung von Daten(quellen) durch schulische Akteur*innen und der darauf beruhenden Innovation von Schule und Unterricht wird ein erhebliches Potenzial für die Adressierung aktueller Herausforderungen des Schulsystems im Sinne verbesserter (Unterrichts-)Prozesse zugeschrieben (Tempelaar et al., 2015).

Jedoch garantiert die gute technische Verfügbarkeit von Daten in einer digitalen (Bildungs-)Welt und deren potenzielle Informativität für die Unterrichtsgestaltung und -entwicklung allein noch nicht deren gewinnbringende Nutzung (Schildkamp, 2019, S. 266). Dies zeigt sich etwa bei bisherigen (analogen) Formen der Datenrückmeldung im Zuge der sogenannten Neuen Steuerung (Altrichter & Maag Merki, 2016): Bisherige empirische Befunde z.B. zu Vergleichsarbeiten belegen bislang kaum die intendierten positiven Wirkungen hinsichtlich (fachlicher) Leistungen von Schüler*innen (Dedering, 2011; Hellrung & Hartig, 2013; Wurster et al., 2017). Des Weiteren stellt sich die Frage nach möglichen Dysfunktionalitäten (Mandinach & Schildkamp, 2021a; Schildkamp, 2019). Dabei spielen Lehrpersonen als Akteursgruppe in diesem Kontext eine zentrale Rolle, da es zu ihren genuinen Aufgaben gehört, Schule und Unterricht zu gestalten und weiterzuentwickeln.

Diese Überlegungen führen zur zentralen These dieses konzeptuellen Beitrags: Um die Potenziale für die datenbasierte Unterrichtsgestaltung und -entwicklung, die mit der Digitalisierung im Bildungswesen einhergehen, ausschöpfen und potenzielle dysfunktionale Wirkungen minimieren zu können, ist die *data literacy* der Lehrkräfte eine notwendige Voraussetzung. Der Beitrag fokussiert dabei insbesondere das Potenzial der Digitalisierung für die individuelle Förderung fachlichen Lernens im Unterricht. Denn das Konzept der individuellen Förderung kann sowohl in der bildungspolitischen Diskussion als auch in der bildungswissenschaftlichen Forschung als ein breit diskutiertes aktuelles Konzept gelten (Dumont, 2019; Hasselhorn et al., 2019). Zudem wird derzeit das Potenzial der Digitalisierung für eine verstärkte Realisierung adaptiver Lehr-Lern-Settings und individuellen Lernens im Bildungskontext von verschiedener Seite betont (Eickelmann, 2018, S. 17; KMK, 2016a; Mandinach & Schildkamp, 2021a; Tempelaar et al., 2015).

Im Beitrag werden zunächst einige Begriffsklärungen vorgenommen und zentrale Modelle des *data-based decision making* skizziert. Dann wird das Konstrukt *data literacy* definiert, von verwandten Konstrukten abgegrenzt und es werden anhand des Forschungsstandes zu *data literacy* die notwendigen Voraussetzungen von Lehrpersonen für das Gelingen von datenbasierter Unterrichtsentwicklung und -gestaltung dargestellt. Anschließend wird das Potential aktueller digitaler Innovationen (technologiebasiertes formatives Assessment und Dashboards) für eine verstärkte Realisierung individueller Förderung analysiert und illustriert, welche zentrale Rolle die *data literacy* der Lehrpersonen für deren Gelingen spielt. Den Beitrag beschließen ein Ausblick auf Forschungsperspektiven sowie die Formulierung von Implikationen für die Praxis.

8.2. Analyse zentraler Modelle aus dem Bereich *data-based decision making*

Data-based decision making (DBDM) kann definiert werden als die systematische Analyse bestehender Datenquellen innerhalb einer Schule, um die Ergebnisse der Analyse anschließend für die Innovation des Unterrichts und des Curriculums sowie der Förderung von Schulleistungen zu verwenden, und diese Innovationen (z.B. allgemeine Entwicklungsmaßnahmen) zu implementieren und weiterzuentwickeln (vgl. Schildkamp & Kuiper, 2010, S. 482). Dabei werden *data-informed decision making, data-driven decision making* und *data use* meist synonym verwendet (Hamilton et al., 2009a; Mandinach & Schildkamp, 2021a; Schildkamp & Kuiper, 2010); im Deutschen wird von daten-basierter/-gestützter Schul- bzw. Unterrichtsentwicklung gesprochen (Altrichter et al., 2016; Wurster, 2019; Wurster et al., 2017). Im Kern geht es also um die systematische Sammlung und Analyse verschiedener Arten von Daten, die für die Entscheidungsfindung im schulischen Kontext herangezogen werden (Hamilton et al., 2009a, S. 46). Dabei wird unterschieden zwischen einem (primären) Fokus auf Rechenschaftslegung (*accountability*), Schulentwicklung (*school improvement*) und Unterricht/-sentwicklung (*instruction/instructional decisions*), wobei diese durchaus in einem Spannungsverhältnis stehen können (Schildkamp, 2019; Schildkamp et al., 2017). Zudem wird konstatiert, dass sich in den letzten Jahren der Fokus von DBDM im Sinne einer ausgeprägteren formativen und kontextbezogenen Perspektive verstärkt hat (Mandinach & Schildkamp, 2021a; van der Kleij et al., 2015). Dabei werden “Daten” als Begriff eher breit verstanden, sodass qualitative, quantitative, formelle Daten wie Lernstandserhebungen und informelle Daten wie Beobachtungen im Klassenzimmer sowie Big Data eingeschlossen sind (Mandinach & Schildkamp, 2021a; Schildkamp, 2019).

Neben Wirkmodellen (z.B. Visscher & Coe, 2003) liegen auch Prozessmodelle vor, die den DBDM-Prozess zu erfassen suchen. Zwei einschlägige Modelle sind hier die von Marsh

(2012) und Schildkamp (2019). Marshs Modell wird wegen seiner Differenziertheit und seiner häufigen Zitierung (vgl. z.B. Kippers et al., 2018; Mandinach & Schildkamp, 2021a; Merk et al., 2020; Reeves, 2017; Schildkamp, 2019) analysiert; das Modell von Schildkamp (2019) wurde aufgrund seiner Aktualität ausgewählt.

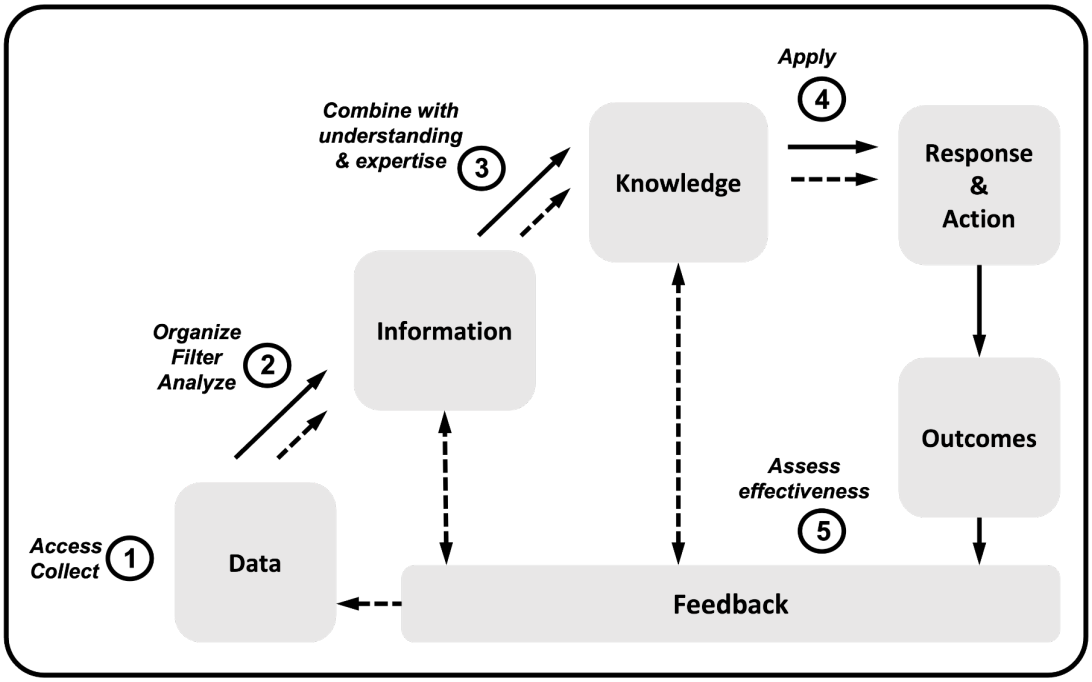


Abbildung 19. DBDM-Prozessmodell nach Marsh

Marshs Modell (2012, vgl. Abbildung 19) basiert auf einem umfangreichen Review und enthält verschiedene Schritte in einem zyklischen Prozess. Zuerst müssen Daten vorhanden sein bzw. gesammelt werden (1), woraus im nächsten Schritt durch Filtern und Analysieren Informationen generiert werden können (2). Indem diese Informationen verstanden und mit Expertise in Verbindung gebracht werden, entsteht Wissen (3), das durch entsprechende Maßnahmen (4) angewendet werden kann, woraufhin die Wirksamkeit dieser Handlungen untersucht werden kann (5). Die gestrichelten Pfeile im Modell deuten an, dass auch innerhalb des zyklischen Prozesses Feedbackschleifen entstehen und neue Daten erhoben werden können.

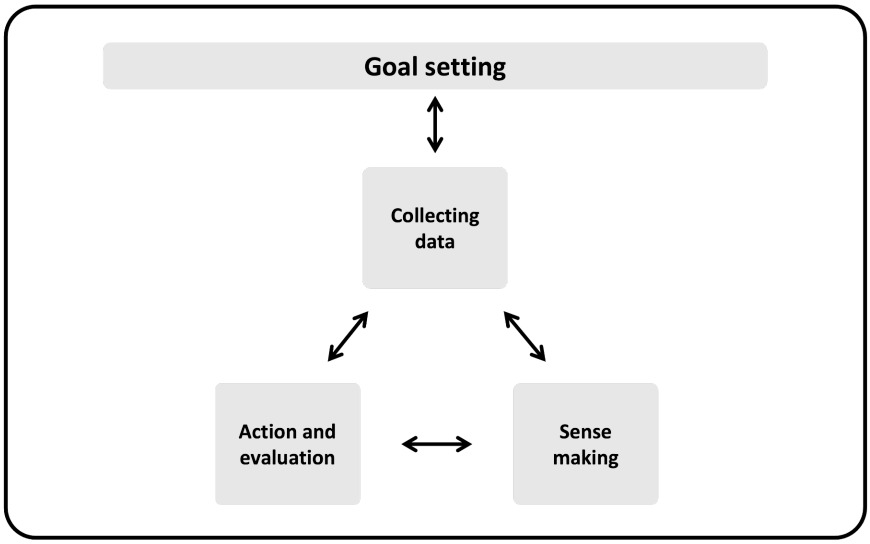


Abbildung 20. DBDM-Modell nach Schildkamp

Das Modell von Schildkamp (2019, vgl. Abbildung 20) unterstreicht in Ergänzung zu dem Modell von Marsh (2012) die Wichtigkeit der Setzung eines klar definierten, spezifischen und messbaren (normativ gesetzten) Ziels zu Beginn des Zyklus, an dem sich alle anderen Schritte orientieren. Auf die Datenerhebung folgt der Schritt des *sense-making*, bei dem die Daten analysiert und interpretiert werden, worauf ein Handlungsplan umgesetzt und evaluiert werden kann (ggf. kann zum vorherigen Schritt zurückgegangen werden, um z.B. ergänzende Daten zu erheben).

Zusammenfassend konzeptualisieren beide Modelle DBDM als einen zyklischen und iterativen Prozess, in dem verschiedene Schritte sequenziell durchlaufen werden. Daraus folgt, dass jeder Schritt auf der Basis des vorhergegangenen Schritts erfolgt und somit eine notwendige Voraussetzung für den darauffolgenden Schritt darstellt. Neben organisationalen Faktoren sind dabei für das Gelingen dieses *data use*-Zyklus auf Ebene der Lehrpersonen ihre Expertise von großem Einfluss, wobei hier die *data literacy* als eine sehr entscheidende Komponente gelten kann (Schildkamp et al., 2017).

8.3. Begriffs- und Konstruktklärung *data literacy*

8.3.1. Definition und Abgrenzung

Mandinach und Gummer (2016) legten im Jahr 2016 eine detaillierte theoretische Konzeptualisierung des Konstrukts *data literacy for teachers* (DLFT) vor, die sie basierend auf umfangreichen Literatursichtungen, eigenen Vorarbeiten, Dokumentanalysen und Expert*inneninterviews entwickelten und die als vergleichsweise umfassend gelten kann (Beck et al., 2019). *Data literacy* wird definiert als

> *"the ability to transform information into actionable instructional knowledge and practices by collecting, analyzing, and interpreting all types of data (assessment, school climate, behavioral, snapshot, longitudinal, moment-to-moment, etc.) to help determine instructional steps. It combines an understanding of data with standards, disciplinary knowledge and practices, curricular knowledge, pedagogical content knowledge, and an understanding of how children learn."* (Mandinach und Gummer 2016, S. 367)

In dieser Konzeptualisierung werden fünf Komponenten unterschieden, die *data literacy* konstituieren (s. Abschnitt 8.3.2), und die Verbindung zu bestimmten Wissensdomänen betont.

Ein verwandtes Konstrukt stellt *assessment literacy* dar, welche ein Verständnis von psychometrischen Gütekriterien und verschiedenen Assessmentformen, die Kommunikation und Nutzung der Ergebnisse für Entscheidungen sowie die Adressierung ethischer Fragen beinhaltet (Beck et al., 2019, S. 8). van der Kleij et al. (2015) sehen in formativem Assessment ein übergeordnetes Konstrukt, in das DBDM (und somit auch *data literacy*) eingegliedert wird. Allerdings bleibt so unberücksichtigt, dass *data literacy* einen summativen Fokus nicht ausschließt und sich zudem auf eine weitaus breitere Datengrundlage als Leistungsdaten beziehen kann, weshalb der Vorschlag von Beck et al. (2019) plausibler erscheint, *data literacy* als übergeordnetes Konstrukt zu sehen, das *assessment literacy* enthält. Beck et al. konstatieren in ihrem systematischen Review allerdings auch, dass Instrumente und Interventionen, die eigentlich *data literacy* adressieren, teils inhaltlich zu *assessment literacy,* allerdings eher bezogen auf nationale Tests und nicht allgemein, tendieren (Beck et al., 2019, S. 27). Ein weiterer verwandter Ansatz zu *data literacy* stellt der Ansatz einer spezifischen *statistical literacy* (Chick & Pierce, 2013; Pierce et al., 2014) dar, der stark auf Kompetenzen für das Verstehen von Ergebnissen nationaler Tests fokussiert und weniger auf statistische Kompetenzen im Allgemeinen. Das statistische Verstehen von (grafisch visualisierten) Testergebnissen wird dabei hierarchisch konzeptualisiert vom Ablesen einzelner Werte bis hin zur Gesamterfassung aller encodierten Entitäten unter Rückbindung an spezifische Kontexte.

Die diagnostische Kompetenz von Lehrpersonen ist ein weiteres adjazentes Konstrukt. Konzeptuelle Gemeinsamkeiten zu *data literacy* in aktuellen Modellen liegen zunächst in der prozesshaften Konzeption der Diagnose (von der Problemidentifikation/ Hypothesengenerierung über die Interpretation von Daten hin zum Ziehen von Schlussfolgerungen), dem Datenbezug sowie der engen Verknüpfung mit (Aspekten von) pädagogischem, fachlichem und fachdidaktischem Wissen bzw. der Konzeption dieser Wissensformen als notwendiger Basis (Heitzmann et al., 2019; Herppich et al., 2018).

Diagnosekompetenz bezieht sich jedoch eher auf die Individualebene von Lernenden sowie lernrelevante Merkmale (Herppich et al., 2018, S. 183), während DBDM/*data literacy* auch die Klassen- und Schulebene adressiert und der Fokus auch nur mittelbar auf Lernen liegen kann (vgl. Abschnitt 8.2 und 8.3). Weiterhin enden Modelle der Diagnostik tendenziell mit dem Stellen und ggf. der Kommunikation der Diagnose (Heitzmann et al., 2019; Herppich et al., 2018), während *data literacy* umfassender und explizit das Treffen einer Handlungsentscheidung sowie deren Evaluation im Prozess abbildet (vgl. Abschnitt 8.3.2).

8.3.2. Komponenten der *data literacy*

Nach Mandinach und Gummer (2016, S. 369–372) besteht *data literacy* aus fünf Komponenten, die jeweils noch einmal in Subfacetten aufgeschlüsselt werden können.

Identifizierung eines Problems/Formulierung von Fragen

Diese Subkomponente (*identify problems and frame questions)* beinhaltet, ein Problem, das z.B. das Curriculum oder einen Aspekt des Unterrichts betreffen kann, zu formulieren. Dies bedingt auch, den Kontext (z.B. die spezifischen Bedingungen der Einzelschule) zu berücksichtigen, ethische Fragestellungen z.B. bezogen auf den Datenschutz der Lernenden zu beachten und ggf. andere Personen miteinzubeziehen.

Datennutzung

Use data bildet die größte Komponente und bezieht sich auf das notwendige Wissen und die Kompetenzen für die direkte Nutzung von Daten. Dazu gehört, dass Lehrpersonen eine oder ggf. mehrere geeignete Datenquelle(n)/-form(en) bezogen auf ihre Fragestellung festlegen, (ggf.) Daten erheben und aufbereiten. Weiterhin geht es um die Analyse der Daten unter Berücksichtigung der Qualität, wozu sowohl eine Aggregierung der Daten zu einer Gesamtaussage als auch eine Disaggregierung, z.B. die Analyse von Fehlkonzepten der Lernenden auf Aufgabenebene, gehören kann.

Transformation der Daten in Information

Diese Komponente (*transform data into information*) zielt auf die Interpretation der Daten unter Berücksichtigung von Kontextinformationen (z.B. aus weiteren Informationsquellen wie dem Unterricht). Hierzu müssen die Daten reflektiert werden, d.h. etwa Muster und Trends in den Daten erkannt, Begründungen/Erklärungen formuliert und Schlüsse gezogen werden.

Transformation der Information in eine Entscheidung

Diese Komponente (*transform information into a decision*) beinhaltet die Konstruktion kontextsensitiver pädagogischer Handlungsmaßnahmen unter Berücksichtigung der

Bedürfnisse der Lernenden. Außerdem zielt sie darauf ab, diese nächsten Schritte umzusetzen, zu überwachen und ggf. Anpassungen vorzunehmen.

Evaluierung der Ergebnisse

Die letzte Subkomponente (*evaluate outcomes*) bezieht sich auf die Evaluation der Effekte der umgesetzten Maßnahmen hinsichtlich des ursprünglich formulierten Problems z.B. im Unterricht oder bezüglich des Verhaltens von Lernenden. Hier wird evaluiert, inwiefern die getroffenen Maßnahmen überhaupt Wirkung bzw. (un)erwünschte Wirkungen zeigen. Ferner geht es darum, in einen neuen DBDM-Zyklus überzugehen.

Rahmende Faktoren

Zusätzlich zu diesen fünf Subkompetenzen der *data literacy* werden in diesem Modell weitere rahmende Faktoren genannt, denen Mandinach und Gummer (2016, S. 372) Bedeutung für das Gelingen von datengestützten Entscheidungen zuschreiben und die mit den Subkompetenzen interagieren, auch wenn sie nicht Teil der eigentlichen *data literacy* sind, weil sie sich stärker auf effektives Unterrichten generell beziehen. Dabei handelt es sich zunächst um domänenspezifisches Wissen (Fachwissen, fachdidaktisches Wissen, allgemeines pädagogisches Wissen, curriculumbezogenes Wissen, Wissen zu Merkmalen von Lernenden und Bildungskontexten/Bildungszielen). Fachliches und fachdidaktisches Wissen bspw. scheinen aus theoretischer Sicht nötig, um sinnvolle Handlungsmaßnahmen auf Basis der Daten konstruieren zu können (Subkomponente Transformation der Information in eine Entscheidung). Weitere Dispositionen bzw. *habits of mind* beinhalten einerseits die Überzeugung, dass alle Schüler*innen lernfähig sind, Schul- bzw. (die eigene) Unterrichtsentwicklung ein beständiger iterativ-zyklischer Prozess ist und datenbasierte Entscheidungen hierbei wirksam sind. Andererseits beziehen sie sich auf ein ethisches Bewusstsein und verantwortungsvolles Verhalten hinsichtlich des Schutzes der persönlichen Daten der Schüler*innen, die Zusammenarbeit mit anderen und auf entsprechende kommunikative Fähigkeiten, um bspw. die Ergebnisse mit verschiedenen Akteur*innen diskutieren zu können. Bei diesen rahmenden Faktoren, die nicht Teil der eigentlichen *data literacy* sind, aber damit eng verbunden postuliert werden, kann kritisch angemerkt werden, dass diese zum einen auf einer konzeptuellen Ebene unverbunden nebeneinander stehen und zum anderen nicht näher spezifiziert bzw. zu anderen Konstrukten abgegrenzt werden (z.B. der Unterschied zwischen *dispositions*, *habits of mind* und *beliefs*/Überzeugungen sowie der Bezug zu Konzeptualisierungen von fachlichem, (fach)didaktischem sowie pädagogischem Wissen und zur diagnostischen Kompetenz).

Insgesamt handelt es sich bei dem Ansatz zu *data literacy* von Mandinach und Gummer (2016) um eine umfassende theoretische Konzeptualisierung der Voraussetzungen bei Lehrpersonen für datengestützte Entscheidungen mit dem Schwerpunkt der Unterrichtsgestaltung und -entwicklung. Es kann herausgestellt werden, dass die fünf Komponenten stark mit den verschiedenen Schritten aus den dargestellten DBDM-Modellen korrespondieren (vgl. Abschnitt 8.2), wobei die Subkomponente Datennutzung als Schlüsselkomponente identifiziert wird (Mandinach & Gummer, 2016, S. 369–372).

8.4. Forschungsstand zu *data literacy*

Nach dieser theoretischen Konzeptualisierung wird, nach einem kurzen illustrierenden Beispiel zur Relevanz von DBDM/*data literacy* im Zuge der Digitalisierung, der empirische Forschungsstand zu *data literacy* und weiteren empirisch identifizierten Einflussfaktoren auf das Gelingen von DBDM auf Ebene der Lehrpersonen ausgehend von aktuellen Überblicksarbeiten skizziert.

Die Relevanz von *data literacy* wird etwa an folgendem illustrierendem Beispiel augenscheinlich: Mithilfe einfach zugänglicher digitaler und endgerätunabhängiger Portale für Feedback zur Unterrichtsqualität von Schüler*innen und/oder im Rahmen kollegialer Hospitation, denen valide Instrumente zugrunde liegen, wie bspw. bei FeedbackSchule (Wisniewski et al., 2020), können Lehrpersonen vergleichsweise einfach valide Rückmeldungen zu von ihnen ausgewählten Aspekten der Unterrichtsqualität bekommen (Identifizierung eines Problems). Vorstellbar sind Systeme, bei denen Rückmeldungen in Echtzeit verfügbar wären und nicht mühsam händisch analog ausgewertet werden müssten. Dann könnten die Einschätzungen zum Unterricht direkt z.B. gemeinsam mit der Klasse exploriert, interpretiert und in handlungsleitende Schlussfolgerungen transformiert werden (Datennutzung, Transformation in Information und Entscheidung). Diese könnten wiederum durch wiederholtes Feedback evaluiert werden (Evaluation), wodurch sich potenziell Entwicklungen in der Unterrichtsqualität von Lehrpersonen abbilden lassen (Bijlsma et al., 2019). Hierin liegt ein ausgeprägtes Potenzial für die Entwicklung von Unterrichtsqualität und damit auch für positive Effekte auf Lernende: Voraussetzung hierfür scheint aber, dass die Daten nicht nur erhoben oder für Zwecke der Rechenschaftslegung genutzt sondern auf Basis einer hinreichend ausgeprägten *data literacy* funktional eingesetzt werden (Rollett et al., 2021).

8.4.1. Empirische Ergebnisse zu den Subkompetenzen von *data literacy*

Identifizierung eines Problems/Formulierung von Fragen

Aus konzeptueller Sicht sind vielfältige Fragestellungen denkbar, die am Anfang eines DBDM-Zyklus von Lehrpersonen formuliert werden können (vgl. Abschnitt 8.3). Studien zeigen, dass in *high-stakes*-Systemen eine starke Fokussierung auf Rechenschaftslegung dazu führen kann, dass Lehrpersonen sich stark auf das Ziel verbesserter Leistungen ihrer Schüler*innen im Sinne von Testscores konzentrieren (Mandinach & Schildkamp, 2021a): Dies kann in Dysfunktionalitäten wie einer Verengung des Curriculums (Au, 2007), *teaching to the test* (Hamilton et al., 2009b), Praktiken des *gaming the system* und die Exkludierung von als schwächer eingeschätzten Lernenden resultieren (Booher-Jennings, 2005; Ehren & Swanborn, 2012). Des Weiteren zeigen Interventionsstudien, dass es Lehrpersonen nicht ohne weiteres gelingt, klar definierte Fragestellungen und Ziele zu formulieren, dies jedoch gefördert werden kann (z.B. Kippers et al., 2018). Insgesamt scheint eine (klare) Zielformulierung für DBDM förderlich zu sein (vgl. Forschungsüberblick bei Schildkamp et al., 2020).

Datennutzung

Sich für Datenquellen zu entscheiden, die sinnvoll auf die Zielsetzung bezogen sind, und diese Daten entsprechend zu erheben und ggf. mehrere Datenquellen zu triangulieren, erscheint für Lehrpersonen anspruchsvoll zu sein (Kippers et al., 2018). Untersuchungen zur Datenanalyse von Lehrpersonen, z.B. mithilfe von Think-Aloud-Interviews, berichten relativ übereinstimmend, dass Daten tendenziell eher unvollständig, oberflächlich sowie auf niedrigen bis mittleren Niveaustufen rezipiert werden (Bez et al., 2021; Gelderblom et al., 2016; Koch, 2011; Means et al., 2011; van der Kleij & Eggen, 2013). In *data literacy*-Tests, die auf diese Subkompetenz zielen, schneiden Lehrpersonen aus dem MINT-Bereich teilweise besser ab (Merk et al., 2020; Zeuch et al., 2017).

Transformation der Daten in Information

Bei dieser Subkompetenz spielt das Vorhandensein von Kontextwissen in breitem Sinne eine Rolle, ohne das die Daten nicht interpretiert werden können, etwa Kontextwissen zu einer bestimmten Kompetenzskala oder auch den spezifischen Bedingungen einer Lerngruppe oder eines lokalen Kontexts (Chick & Pierce, 2013; Pierce et al., 2014). Daneben heben bezüglich dieser Subkomponente verschiedene Überblicksarbeiten die Rolle von Vorannahmen wie bspw. der eigene Eindruck von eine*r Schüler*in oder einer Klasse sowie Überzeugungen von Lehrkräften hervor (Coburn & Turner, 2011; Mandinach & Schildkamp, 2021; Schildkamp, 2019): Vanlommel et al. bspw. fanden in mehreren Studien zu Übergangsentscheidungen verschiedene Urteilsverzerrungen, die darauf basierten, dass bei der Informationsgenerierung

Überzeugungen und Vorannahmen stärker gewichtet wurden als die Daten (Vanlommel et al., 2017, 2020; Vanlommel & Schildkamp, 2019). Vorannahmen, die durch Daten bestätigt werden, können dabei auch zur Verstärkung bestimmter Erwartungen bezogen auf z.B. als schwächer eingeschätzte Schüler*innen auf Grundlage bestimmter Merkmale fungieren, wie in verschiedenen explorativen Studien beobachtet bzw. dokumentiert werden konnte (z.B. Bertrand & Marsh, 2015; Datnow & Park, 2018).

Transformation der Information in eine Entscheidung

Für das Treffen von konkreten kontextsensitiven Handlungsmaßnahmen auf Basis der Datenanalyse und -interpretation scheint es notwendig, vorhandenes Fach- und (fach-)didaktisches Wissen adäquat mit den Daten z.B. für Differenzierungsmaßnahmen zu verknüpfen, was Lehrpersonen nicht ohne weiteres leicht zu fallen scheint (Kippers et al., 2018; Mandinach & Schildkamp, 2021a; Schildkamp, 2019).

Evaluierung der Ergebnisse

Übergreifend wird betont, dass konzeptuelle Veränderungen nicht notwendigerweise zu konkreten Maßnahmen, also zur direkten Nutzung der Daten im Sinne einer Verbesserung und Entwicklung, führen müssen (Altrichter et al., 2016). So können auch indirekte, symbolische und strategische Maßnahmen getroffen, umgesetzt und evaluiert werden, die nicht notwendigerweise (grundlegende) Veränderungen des unterrichtlichen Handelns nach sich ziehen müssen (Datnow & Park, 2018; Groß Ophoff, 2013b; Schildkamp, 2019), sondern sich subtil bis offen dysfunktional bzw. nachteilig für Schüler*innen auswirken können (s.o. Ausführungen zur Problemidentifizierung).

8.4.2. Sonstige Voraussetzungen

Neben *data literacy* werden weitere Voraussetzungen für das Gelingen von DBDM berichtet. Dazu gehören auf personaler Ebene zunächst Überzeugungen und Einstellungen, wobei positive Überzeugungen bzw. Einstellungen zur Wirkung und Nützlichkeit von datengestützten Entscheidungen als förderlich gelten (Datnow & Hubbard, 2016; Heitink et al., 2016; Hoogland et al., 2016; Schildkamp et al., 2020). Des Weiteren werden die wahrgenommene Selbstwirksamkeit, Kontrolle und sozialer Druck sowie kollegiale Kooperation aufgeführt, wobei ein gewisses Maß an sozialer Verbindlichkeit als förderlich gelten und eine ausgeprägte Wahrnehmung von äußerer Kontrolle als eher einschränkend gelten kann (Dunn et al., 2013; Heitink et al., 2016; Schildkamp et al., 2020).

Als weitere Voraussetzung auf kognitiver Ebene wird zunächst die Notwendigkeit von fachlichem, (fach-)didaktischem und pädagogischem Wissen betont (Hoogland et al., 2016;

Reeves, 2017; Schildkamp et al., 2020), da Daten dazu beitragen können, z.B. bestimmte Fehlkonzepte offen zu legen. Dies kann als Voraussetzung für die Planung und (differenzierte) Umsetzung nachfolgender Lehr-Lern-Schritte gelten. Weiterhin werden digitalisierungsbezogene Fertigkeiten als Faktor dafür genannt, dass Lehrpersonen z.B. ein schulinternes System oder bestimmte Plattformen/Tools adäquat für DBDM einsetzen können (Hoogland et al., 2016; Schildkamp et al., 2020).

8.4.3. Zwischenfazit

Bisher wurde im Beitrag ein Überblick zu theoretischen Konzeptualisierungen und empirischen Forschungsergebnissen zu DBDM und *data literacy* gegeben. Insgesamt gelten die oben ausgeführten Subkompetenzen von *data literacy* als zentrale Voraussetzungen für das Gelingen von DBDM. Bezogen auf positive Effekte von DBDM, wobei hier in der Regel die (fachlichen) Leistungen von Schüler*innen fokussiert werden, ergibt sich national wie international ein gemischtes Bild: Studien, die über retrospektive Selbstauskünfte hinausgehen und (quasi-)experimentell angelegt sind, zeigen teilweise positive Effekte, aber bei weitem nicht durchgehend (Altrichter et al., 2016; Marsh, 2012; Visscher, 2021). Im Überblick von Visscher (2021) bezogen auf sechs (quasi-)experimentelle Interventionsstudien in den Niederlanden zeigten vier Studien signifikante positive und drei davon differenzielle Effekte auf fachliche Leistungen (bei Schüler*innen in Klasse 3-8 in verschiedenen Fächern gemessen in standardisierten Tests), wobei die Lernzuwächse bei schwächer abschneidenden Schüler*innen und Lernenden sowohl mit niedrigem als auch mit hohem sozioökonomischen Hintergrund größer waren. Dabei wird die Entwicklung entsprechender Kompetenzen bezogen auf den gesamten Zyklus von DBDM als ein besonders wichtiger Faktor dafür herausgestellt, dass sich das unterrichtliche Handeln der Lehrpersonen z.B. bezüglich verstärkter adäquater Differenzierungsmaßnahmen wirklich verändert und in entsprechenden positiven Veränderungen bei den Lernenden niederschlagen kann (Visscher, 2021).

Somit kann anhand sowohl theoretischer Modelle als auch empirischer Forschungsergebnisse abgeleitet werden, dass *data literacy* eine notwendige Voraussetzung für das Verhindern von Dysfunktionalitäten (vgl. Abschnitt 8.4.1 und 8.5) und das Gelingen des komplexen Prozesses von datengestützten Entscheidungen im schulischen Kontext im Sinne verbesserter Lernprozesse von Schüler*innen darstellt. Dies wird nun ins Verhältnis zu aktuellen digitalen Innovationen mit Bezug zur verstärkten individuellen Förderung von Lernenden gesetzt.

8.5. Illustration für die Notwendigkeit hinreichender *data literacy* anhand ausgewählter aktueller Innovationen

8.5.1. Technologiebasiertes formatives Assessment

Technologiebasierte formative Assessmentsysteme bieten die Möglichkeit, im Unterricht regelmäßig mit *allen* Lernenden einer Klasse konstruktvalide, auf kalibrierten Aufgabenpools aufbauende Lernstands- bzw. Lernverlaufsmessungen durchzuführen, was aufgrund des zu leistenden Aufwandes mit analogen, selbstkonstruierten Assessments durch Lehrpersonen in dieser Validität in der alltäglichen Praxis kaum möglich wäre (Schütze et al., 2018, S. 710; Souvignier et al., 2014, S. 241). Des Weiteren erlauben digitale Systeme aufgrund sophistizierter psychometrischer Modelle überhaupt erst adaptives Testen sowie Vergleiche von Lernständen über die Zeit als auch zwischen Gruppen (Conole & Warburton, 2005; Tomasik et al., 2018). *Potenziell* verfügen Lehrpersonen damit über ein digitales Tool, das ihnen ermöglicht, auf Basis qualitativ hochwertiger Lernstands- und Lernverlaufsmessungen datenbasierte Entscheidungen für ihren Unterricht zu treffen, hochgradig adaptiv zu unterrichten bzw. Lernende individuell zu fördern und die getroffenen Entscheidungen anhand erneuter valider Messungen zu evaluieren und weiterzuentwickeln. Aus konzeptueller Sicht hängen die Entfaltung der skizzierten Vorteile und die Minimierung möglicher dysfunktionaler Wirkungen dieses `digitalen´ DBDM in der Praxis jedoch in hohem Maß von der *data literacy* der Lehrpersonen ab: Sind Lehrpersonen z.B. nicht in der Lage, ein Problem zu identifizieren bzw. eine formative Perspektive bei der Formulierung von Fragen einzunehmen, führt das beim Einsatz solcher Assessmentsysteme möglicherweise dazu, dass sie mit dem Fokus auf Rechenschaftslegung oder (mehr oder weniger subtil) als Selektionsdiagnostik eingesetzt werden. Oder rezipieren Lehrpersonen die Ergebnisse ihrer Klasse, d.h. visualisierte Lernstände und Lernverläufe nur oberflächlich oder inadäquat, besteht die Gefahr, dass potenziell wichtige Informationen nicht generiert werden: Durch einen vorrangigen Fokus auf den Klassenmittelwert oder die soziale Bezugsnorm etwa können Lehrpersonen individuelle Stagnationen oder substanzielle Lernfortschritte übersehen. Damit bliebe die eigentliche Stärke einer exakten Erfassung von Lernverläufen sowie die Möglichkeit einer ipsativen Bezugsnorm, die ja überhaupt erst durch dieses digitale Tool ermöglicht werden, unberücksichtigt. Dies wäre aber die Voraussetzung für die adaptive kontextsensitive Gestaltung von zukünftigen Lehr-Lernprozessen mit einer einhergehenden verbesserten Förderung aller Lernenden und der fortlaufenden Evaluation des DBDM-Prozesses.

8.5.2. Dashboards

Dashboards sind Displays, die Indikatoren über Lernende, Lernprozesse und/oder Lernkontexte aggregiert in einer (oder mehreren) Visualisierung(en) darstellen (vgl. Schwendimann et al., 2017, S. 37). So können etwa in Echtzeit die Bearbeitungsdauer einer

Aufgabenserie, Kontextmerkmale sowie Lernergebnisse wie die Anzahl korrekt bearbeiteter Aufgaben von Schüler*innen z. B. auf Tablets dargestellt werden (Molenaar & Knoop-van Campen, 2019). Dashboards gelten -nach dieser Definition- als eine Anwendung von Learning Analytics (Molenaar & Knoop-van Campen, 2019; Verbert et al., 2013), wobei letztere zum Ziel haben, durch die Messung, Sammlung, Analyse und Kommunikation von Daten über Lernende und ihre Kontexte Lernen und Lernumgebungen besser zu verstehen und zu optimieren (vgl. Siemens & Gašević, 2012, S. 1). So werden neue und breitere Datenquellen erschlossen (Schildkamp, 2019, S. 262–264) und durch Triangulation, Analyse und Aufbereitung auch potenziell bis dahin unbekannte Zusammenhänge und Muster exploriert (Wang, 2021). Dashboards können unterschiedliche Nutzer*innen (Lehrende, Lernende, ...) adressieren und sich zudem in ihren Kontexten, Datengrundlagen und Anwendungen sehr unterscheiden (z.B. *massive open online courses* im Hochschulkontext vs. schulischer Kontext, Prädizierung von Dropout vs. Performanzüberblick, Fokus auf kognitive vs. metakognitive Lernziele usw.) (vgl. aktuelle Reviews von Du et al., 2021; Jivet et al., 2017; Schwendimann et al., 2017). Daher werden zunächst exemplarisch Dashboards für Lehrpersonen in synchronen Unterrichtssettings fokussiert und dann auf einer allgemeineren Ebene die Anforderungen an Lehrpersonen bzgl. *data literacy* bei Dashboards skizziert.

Dashboards im (synchronen) Unterricht wird Potenzial zur Unterstützung von Lehrpersonen zugesprochen, da sie schnell einen besseren Überblick über das Lerngeschehen im Klassenzimmer ermöglichen (z.B. wer/welche Gruppe arbeitet gerade an welcher Stelle im Lernmodul), Einblicke in individuelle Lernprozesse geben (z.B. wer braucht wie lange für eine Aufgabe) und somit auch individuelle Schwierigkeiten sichtbar machen können (z.B. wer konnte Aufgaben nicht lösen, die auf die Aktivierung notwendigen Vorwissens zielen) (van Leeuwen, 2015). Diese Schwierigkeiten können Lehrpersonen dann adaptiv direkt in der Situation aufgreifen und so Lernende individuell oder in Kleingruppen fördern, z.B. durch entsprechendes Feedback (Knoop-van Campen et al., 2021; Knoop-van Campen & Molenaar, 2020; Verbert et al., 2013). Allerdings scheint es für Lehrpersonen durchaus herausfordernd, Dashboards in ihr unterrichtliches Handeln zu integrieren (Keuning & van Geel, 2021; Molenaar & Knoop-van Campen, 2019) und *data literacy* eine notwendige Voraussetzung dafür zu sein, dass sich die Potenziale für eine verstärkte Realisierung von individueller Förderung im Unterricht entfalten können. Explorieren etwa Lehrpersonen im laufenden Unterricht die in Echtzeit generierten Daten nur oberflächlich und erkennen z.B. verschiedene Leistungsgruppen in den Daten nicht oder unzureichend, werden darauffolgende Differenzierungsmaßnahmen wie die Bildung homogener Kleingruppen kaum erfolgreich sein. Des Weiteren müssen Daten aus Dashboards nicht nur analysiert sondern auch interpretiert werden, indem Lehrpersonen etwa reflektieren müssen, ob ein*e Schüler*in aktuell

Schwierigkeiten bei einer Aufgabenbearbeitung hat (und wenn ja, auf welcher Ebene) und Unterstützung benötigt oder schlicht noch etwas Zeit braucht. Eine weitere Herausforderung besteht darin, unter Einbeziehung pädagogischen, fachlichen und fachdidaktischen Wissens auf Grundlage der Daten angemessene pädagogische Handlungsmaßnahmen zu konstruieren, etwa sinnvolle Lerngruppen zu bilden oder Fehlkonzepte einzelner Lernenden durch adäquates Feedback zu adressieren, diese umzusetzen und wiederum in ihrer Wirksamkeit zu evaluieren. Der Aspekt der Echtzeit im Unterricht erhöht die Anforderungen an Lehrpersonen dahingehend, als dass diese Prozesse beschleunigt und in der Simultanität und Unmittelbarkeit des Unterrichts (Doyle, 1986) ablaufen.

Wie bereits zu Beginn des Abschnitts angedeutet, zeigen aktuelle Übersichtsarbeiten (die sich allerdings meist auf Lernende als Nutzer*innen beziehen), dass sich Dashboards in der Intensität ihrer lehr-lerntheoretischen Fundierung, der adressierten Ziele/Kompetenzen und Datengrundlagen deutlich unterscheiden (Du et al., 2021; Jivet et al., 2017, 2018). Des Weiteren arbeiten die Autor*innen die Dominanz der sozialen Bezugsnorm heraus, d.h. Vergleiche mit Daten von Peers, während individuelle und bzw. an einer sachlichen Bezugsnorm orientierte Foki seltener vorkommen; dies wird etwa für Lernende, die wiederholt im sozialen Vergleich schlecht abschneiden (würden), aus motivationspsychologischer Perspektive problematisiert (Jivet et al., 2017, 2018). Jarke und Macgilchrist (2021) zeigen aus einer *storytelling*-Perspektive weitere mögliche dysfunktionale Wirkungen auf Lernende auf: Sie formulieren etwa die Gefahr von (fälschlichen) kausalen Interpretationen von korrelativen Daten auf Seiten der Lehrpersonen sowie die aufgrund einer missverstandenen Objektivität mögliche problematische Individualisierung von eigentlich strukturellen Faktoren. Dies kann zu Verzerrungen in den Scores bzw. in prädiktiven Systemen und bei mangelnder Berücksichtigung potenziell zu dysfunktionalen Entscheidungen durch Lehrpersonen und daraus folgende problematischen Konsequenzen z.B. im Sinne einer verstärkten Marginalisierung für Lernende mit bestimmten Merkmalen führen (Jarke & Macgilchrist, 2021).

Mit Blick auf *data literacy* lässt sich daraus ableiten, dass die Wahrscheinlichkeit eines dysfunktionalen Einsatzes solcher Dashboards durch Lehrpersonen sinkt, wenn sie über hinreichende *data literacy* verfügen. Denn dann kann angenommen werden, dass Lehrpersonen z.B. Dashboards in Passung zu spezifischen Fragestellungen einsetzen und hinsichtlich der Referenzrahmen und lehr-lerntheoretischen Fundierung kritisch evaluieren. Wenn weiterhin Lehrpersonen z.B. basale Kenntnisse zu zugrundeliegenden Mechanismen von Dashbords besitzen und mögliche Urteilsverzerrungen reflektieren können, sinkt die Gefahr von Über- bzw. Falschinterpretationen der (visualisierten) Daten und darauf basierenden dysfunktionalen Entscheidungen für Lernende. Insgesamt scheint hier jedoch,

u.a. aufgrund des noch relativ jungen Feldes zu Learning Analytics Dashboards, ein großer diesbezüglicher Forschungsbedarf zu bestehen (Jude et al., 2020).

8.6. Fazit und Perspektiven

Im Beitrag wurden Potenziale der Digitalisierung für datengestützte Entscheidungen (DBDM) auf der Ebene von Lehrpersonen fokussiert. Dabei wurde zentral anhand theoretischer Konzeptualisierungen, einschlägiger Forschungsergebnisse und der Analyse aktueller Innovationen für die These argumentiert, dass die *data literacy* von Lehrpersonen eine notwendige Voraussetzung für das Gelingen verstärkter individueller Förderung mithilfe digitaler Technologie im Rahmen von DBDM ist. Einschränkend ist allerdings zu reflektieren, dass dieser Beitrag durch seine Konzeption notwendigerweise selektiv (wenn auch begründet) z.B. bei der Auswahl zentraler Modelle ist und etwa hinsichtlich der Darstellung des Forschungsstandes nicht den Kriterien eines systematischen Review Genüge leistet. Des Weiteren werden durch den Fokus auf die Ebene der Lehrpersonen stärker mehrebenenperspektivische und organisationale Faktoren weniger berücksichtigt; so können empirisch gesehen u.a. die Rolle der Schulleitung, gelingende Kooperationen innerhalb der Einzelschule oder eine hinreichende technische Ausstattung als relevant für DBDM gelten (Schildkamp et al., 2017).

Basierend auf den bisherigen Überlegungen werden als Ausblick Perspektiven für die Bildungsforschung und die Bildungspraxis formuliert. Zwar scheint aktuell nur schwer absehbar, welche digitalen Innovationen in welcher Form wann in Deutschland, z.B. aus Datenschutzgründen, flächendeckend implementiert werden, Ansätze zu formativem Assessment/digitaler Lernverlaufsdiagnostik sind jedoch deutlich erkennbar (Jude et al., 2020). Die zunehmende Digitalisierung macht grundsätzlich die einfachere und mannigfaltige Genese von Daten, ihre Aufbereitung, Triangulation, Analyse und Dissemination deutlich erleichtert möglich, was an sich aber noch nicht die potenzielle und vor allem gewinnbringende Nutzung für DBDM hinsichtlich verstärkter individueller Förderung garantiert, sondern gewisse Anforderungen an Lehrkräfte noch verstärkt. Durch die zentrale Stellung der *data literacy* sollte diese stärker in den Fokus von bildungswissenschaftlicher Forschung und Aus-/Weiterbildung der Lehrpersonen gerückt werden. Dabei stellt sich als übergeordneter interdisziplinärer Forschungsbedarf die Frage nach einer gelingenden Interaktion von Mensch und Technologie mit dem Fokus auf pädagogischer Fundierung und pädagogischen Zielen, wie er sowohl von Seiten der DBDM-Forschung als auch der Learning-Analytics-Forschung übereinstimmend formuliert wird (Du et al., 2021; Jivet et al., 2017, 2018; Mandinach & Schildkamp 2021a; Verbert et al., 2020). Eine virulente Frage besteht hier etwa in der Entwicklung von Darstellungen bzw. Visualisierungen von Daten, die eine effiziente und

akkurate Rezeption durch die adressierten Nutzer*innen in Passung zu ihrer individuellen *data literacy* als entscheidenden Schritt im zyklischen Prozess ermöglicht, ohne übersimplifizierend oder überkomplex zu sein (vgl. Diskussion bei Mandinach & Schildkamp, 2021a; Verbert et al., 2020). Diese Frage ist nicht trivial, da einerseits unterschiedliche Visualisierungen derselben Daten unterschiedliche Schlüsse nahelegen können (vgl. hierzu ein aktuelles Beispiel zur Darstellung von Lernverläufen in Ratner et al., 2019, S. 31) und andererseits sich bestimmte Vorteile z.B. hinsichtlich der Validität digitaler Tools nur dann entfalten können, wenn die Ergebnisse nicht fehlanalysiert oder missinterpretiert werden (vgl. Abschnitt 8.5). Eine weitere Forschungsfrage, gerade auch im Kontext von Big Data, besteht darin, welche Arten und Triangulationen von Daten für welche Arten von Fragestellungen und Entscheidungen pädagogisch (ir)relevant bzw. (dys)funktional sind (Schildkamp, 2019, S. 261–264). Außerdem scheint die datenschutz- und persönlichkeitsrechtliche sowie ethische Fragen betreffende kritisch-sensibilisierende Komponente von *data literacy* und deren Förderung auch (aber nicht nur) auf Ebene der Lehrpersonen wichtiger zu werden, um problematische Wirkungen für (in der Regel minderjährige) Lernende zu verhindern: Dies betrifft z.B. den Zugang und die Verknüpfung von Daten (Schildkamp, 2019, S. 263–264), eine kritische Technikfolgenabschätzung und die Reflexion möglicher Urteilsverzerrungen gerade auch beim Einsatz von Learning Analytics und Recommender-Systemen (Eyal, 2012; Hartong, 2019; Jarke & Macgilchrist, 2021; Wang, 2021), um etwa dysfunktionale Folgen bezogen auf einzelne Schüler*innen oder Gruppen von Lernenden mit bestimmten Merkmalen vorzubeugen (Datnow & Park, 2018).

Durch die herausgearbeitete Relevanz von *data literacy* besteht eine zentrale Implikation für die Praxis zunächst in der Adressierung von *data literacy* in der Aus- und Weiterbildung von Lehrpersonen, da Lehrpersonen DBDM insgesamt als herausfordernd wahrnehmen (Mandinach & Schildkamp, 2021a). Hierzu liegen bereits theoretische Konzepte einer kontinuierlichen Anbahnung von *data literacy* über verschiedene Phasen der Lehrer*innenbildung hinweg (z.B. Beck & Nunnaley, 2020) sowie empirisch überprüfte wirksame Interventionen für Lehramtsstudierende und Lehrpersonen (z.B. Merk et al., 2020; van Geel et al., 2017; Visscher, 2021) vor. In der schulischen Praxis scheint insbesondere die Implementation von kollegialer Kooperation und *data teams*, möglicherweise mit externer Unterstützung, vielversprechend (Schildkamp et al., 2016, 2019). Speziell für die Lehramtsausbildung stellt sich die Frage nach einer gewinnbringenden Adressierung von *data literacy* in Verbindung mit fachlichen, fachdidaktischen u. ä. Kompetenzen, da deren Verknüpfung an verschiedenen Stellen im DBDM-Zyklus relevant wird (vgl. Abschnitt 8.3, 8.4 und 8.5). Durch konzeptionelle Ähnlichkeiten zu Subfacetten der diagnostischen Kompetenz (vgl. Abschnitt 8.3) erscheint es außerdem sinnvoll, in den verschiedenen Phasen der

Lehrer*innenbildung zu versuchen, auch diesbezüglich Synergieeffekte zu induzieren. Denn insgesamt beansprucht die *data literacy* von Lehrpersonen für das Gelingen datengestützter Unterrichtsgestaltung und -entwicklung in einer digitalen Bildungswelt entscheidende Bedeutung und bedarf sowohl weiterer Forschung, der Adressierung in der Aus- und Weiterbildung von Lehrpersonen als auch der Implementation gezielter Unterstützung für Lehrpersonen in der Praxis.

9. Diskussion

Das Forschungsinteresse dieser Arbeit bezieht sich auf die Datenrezeption und -interpretation von Lehrpersonen bei datengestützten Entscheidungen, da ihnen im Prozess der datenbasierten Gestaltung und Entwicklung von Schule und Unterricht eine zentrale Rolle zukommt. In diesem Kapitel werden zunächst die zentralen Ergebnisse der Arbeit anhand der drei Forschungsfragen zusammengefasst und diskutiert, bevor zentrale Limitationen der Arbeit formuliert werden. Es folgen Implikationen für sich an diese Arbeit anschließende Forschungsfragen sowie abschließend Implikationen für die schulische Praxis sowie die Aus- und Weiterbildung von Lehrpersonen.

9.1. Zusammenfassung und Diskussion der Ergebnisse

Bezüglich Forschungsfrage 1 (*Wie rezipieren und interpretieren Lehrpersonen Daten bei datengestützten Entscheidungen?*) deuten die durchgeführten explorativen Think-Aloud-Studien darauf hin, dass Lehrpersonen in Grafiken visualisierte Leistungsdaten ihrer Klassen in der Tendenz mit niedriger bis mittlerer Komplexität rezipieren. Das bedeutet, dass sie vorrangig direkt gegebene Entitäten in den Visualisierungen ablesen oder miteinander vergleichen, aber nicht unbedingt die Daten in ihrer Gesamtheit erfassen. Dabei zeigt sich allerdings substanzielle Heterogenität in allen Studien, also sowohl in den deskriptiven Ergebnissen von Artikel 1 und 2 als auch in den Ergebnissen der Clusteranalyse in Artikel 2. Damit reiht sich diese Arbeit in die Ergebnisse anderer (explorativer) Studien ein, die darauf hindeuten, dass Lehrkräfte bei der Rezeption und Interpretation von Daten im Rahmen datenbasierter Entscheidungen die höchste Komplexitätsstufe von *graph literacy* in der Tendenz eher nicht erreichen (Espin et al., 2017; Goffin et al., 2023; van den Bosch et al., 2017; Zeuch et al., 2017). Dieses Ergebnis scheint sich also auch zu zeigen, wenn sich die Datengrundlage auf die eigenen Klassen der Lehrkräfte bezieht, d.h. sie über Kontextwissen und Erfahrungswissen (zu ihren Schüler*innen, dem bisherigen gemeinsamen Unterricht, spezifischen Besonderheiten der Einzelschule, usw.) verfügen. Jenseits dieser globalen Betrachtung tätigen einzelne Lehrkräfte durchaus elaborierte Äußerungen, z.B. was die

Analyse von Fehlern in Aufgabenergebnissen einzelner Schüler*innen anbelangt (vgl. Artikel 2). Insgesamt kann dieses Ergebnis als ein weiterer Hinweis dahingehend interpretiert werden, dass Lehrpersonen in der Tendenz Schwierigkeiten damit haben, (Leistungs-)Daten zu rezipieren, zu interpretieren (Mandinach & Schildkamp, 2021) und dabei die Datengrundlage in ihrer Gesamtheit zu erfassen. Damit unterstreicht dieses Ergebnis auch die Relevanz von Forschung zur Förderung von Datenrezeption und -interpretation bei angehenden Lehrpersonen (Forschungsfrage 2).

Des Weiteren lässt sich aus dem explorierten Prozessmodell in Artikel 2 auf der Basis der kodierten Prozessschritte die Hypothese generieren, dass die Analyse von Fehlern in den Aufgabenergebnissen der Schüler*innen und der Abgleich der Ergebnisse mit der eigenen Einschätzung wichtige Interpretationsschritte von Lehrpersonen für die Konstruktion unterrichtlicher Anschlusshandlungen sind und dafür die alleinige Datenrezeption nicht ausreichend zu sein scheint. Bei der Analyse der Fehler in den Aufgabenergebnissen, z.B. bezüglich typischer und individueller fachlicher Fehlkonzepte, spielt fachliches und fachdidaktisches Wissen explizit eine Rolle. Diese Einbeziehung von fachlichem und fachdidaktischem Wissen korrespondiert eng mit dem Modell und der Definition von *data literacy* nach Mandinach und Gummer (2016), dem die Relationierung zu Domänen professionellen Wissens nach Shulman (1987) inhärent ist.

Aus einer Prozessperspektive zeigen sich als weiteres Ergebnis in den verbalisierten Kognitionen zahlreiche Iterationen und Rekursionen (vgl. Abbildung 9 und 12), während die konzeptuellen Prozessmodelle (vgl. Kapitel 2.2) in der Tendenz linear-sequenziell aufgebaut sind. Wenn sich dies in weiteren Studien zeigt, könnte eine stärkere Berücksichtigung dessen in den konzeptuellen Modellen in Betracht gezogen werden, in dem z.B. der Fokus stärker auf die Verknüpfungen der einzelnen Prozessschritte gelegt werden könnte.

Neben den dargestellten inhaltlichen Ergebnissen bezüglich Forschungsfrage 1 trägt diese Arbeit auch auf der methodischen Ebene dazu bei, größere Alltagsnähe und eine Prozessorientierung auf einer Mikroebene bei datengestützten Entscheidungen von Lehrpersonen zu adressieren, was beides als wichtig erachtet wird, zu dem bisher aber nur wenige Studien vorliegen (Goffin et al., 2022; Hebbecker et al., 2022; Schildkamp, 2019): Anlehnend und ergänzend an bisherige Think-Aloud-Studien in diesem Feld wurden als Datengrundlage eigene Daten von Lehrpersonen bzw. deren Schüler*innen verwendet. Zudem wurden die Aufzeichnungen des lauten Denkens als *timed-event codings* (Bakeman & Quera, 2011) kodiert. Dies ist zwar ressourcenintensiv, ermöglicht aber die Analyse von verbalisierten Kognitionen auf der Prozessebene jenseits der Häufigkeiten von vergebenen

Kategorien (Hartmann et al., 2022; Reimann, 2009; Sonnenberg & Bannert, 2019). Neben der Visualisierung der Verläufe wurde auf der Basis der Kodierungen *process mining* (van der Aalst, 2016) angewandt. Dieser Ansatz aus dem Bereich *data science* kann mithilfe von Algorithmen den typischen Prozess in Daten, die den Start und das Ende von Ereignissen enthalten, explorieren und in einem Modell darstellen. Durch diesen methodischen Ansatz und die Anwendung von neuen Methoden aus dem Bereich *data science* leistet diese Arbeit auch einen methodischen Beitrag dazu, die alltägliche Datenrezeption und -interpretation von Lehrkräften aus einer Prozessperspektive zu adressieren.

Bezüglich Forschungsfrage 2 (*Wie können die Datenrezeption und -interpretation bei angehenden Lehrpersonen gefördert werden?*) zeigen die Ergebnisse der vorliegenden Arbeit im Rahmen von Artikel 2 Evidenz für einen großen positiven Effekt der Intervention auf die Kompetenzen der Lehramtsstudierenden. Somit ist der Effekt der Onlineintervention ähnlich groß wie in der ursprünglichen Präsenzform der Intervention von Merk et al. (2020). Damit trägt diese konzeptuelle Replikationsstudie zu Erkenntnissen bei, wie in einer relativ kurzen Zeit in einer digitalen Lernumgebung die Datenrezeption und -interpretation von angehenden Lehrpersonen für die datengestützte Gestaltung und Entwicklung von Schule und Unterricht gefördert werden können, und geht dabei in methodischer Hinsicht über andere ähnliche Studien, die rein auf Selbstauskünften basieren (z.B. Reeves & Chiang, 2018, 2019; Supovitz & Sirinides, 2018), hinaus. Zudem zeigt die Intervention auch einen positiven Effekt auf Komponenten der motivationalen Überzeugungen von Lehramtsstudierenden bezüglich datengestützter Entscheidungen, insbesondere auf ihre Selbstwirksamkeitserwartung.

Bezüglich Forschungsfrage 3 (*Welche Rolle spielt data literacy im Kontext zunehmender Digitalisierung für datengestützte Entscheidungen?*) wird auf einer konzeptuellen Ebene basierend auf theoretischen Modellen und einschlägigen Forschungsergebnissen für die These argumentiert, dass die *data literacy* von Lehrkräften eine notwendige Voraussetzung für das Gelingen verstärkter individueller Förderung mithilfe digitaler Technologie im Rahmen von datengestützten Entscheidungen ist. Dabei wird exemplarisch anhand aktueller digitaler Innovationen, nämlich technologiebasiertem formativen Assessment und Dashboards im (synchronen) Unterricht, veranschaulicht, welche zentrale Rolle die *data literacy* von Lehrpersonen spielt, damit mögliche dysfunktionale Wirkungen minimiert werden und den Innovationen erst einmal nur zugeschriebene Potenziale sich entfalten können. Damit reiht sich diese Arbeit in andere ähnliche konzeptuelle Arbeiten ein, die z.B. Digitalisierung und die diagnostische Kompetenz von Lehrpersonen aufeinander beziehen (z.B. Gottlebe et al., 2023) oder allgemeinere Rahmenmodelle für technologiegestützte adaptive Lehr-Lernsettings entwickeln (z.B. Kärner et al., 2021). Sie nimmt dabei auch die Forderung nach kritischen

konzeptuellen Arbeiten auf, die sowohl positive als auch negative Seiten beleuchten (Krein & Schiefner-Rohs, 2021), und versucht exemplarisch zu entwickeln, welche Schlüsselrolle Lehrpersonen für die Gestaltung und Entwicklung unterrichtlicher Prozesse mithilfe aktueller digitaler Innovationen im Bereich datengestützter Entscheidungen einnehmen.

9.2. Limitationen

Im Folgenden werden zentrale, übergreifende Limitationen der vorliegenden Arbeit formuliert und diskutiert. Da in den entsprechenden Abschnitten der Diskussionskapitel bereits die Limitationen der einzelnen Studien diskutiert werden, liegt der Fokus hier auf übergreifenden Aspekten.

Auf der Basis der Prozessmodelle und bisheriger Forschungsergebnisse wurde in dieser Arbeit dafür argumentiert, dass die (adäquate) Rezeption und Interpretation von Lehrpersonen im Rahmen datengestützter Entscheidungen zentral sind, weshalb sie in den Fokus dieser Arbeit gestellt wurden. Allerdings sind die Rezeption und Interpretation zwar notwendige aber keinesfalls hinreichende Bedingungen: Nur wenn anschließend (angemessene) Schlussfolgerungen für pädagogisch-didaktische Anschlusshandlungen gezogen und diese auch umgesetzt werden, können sich überhaupt Wirkungen von datengestützten Entscheidungen zeigen. Daher besteht eine zentrale Limitation der Arbeit darin, dass sie nicht die gesamte Wirkungskette datengestützter Entscheidungen und nicht *data literacy* als Gesamtkonstrukt untersucht, sondern sich auf die Teilschritte der Datenrezeption und -interpretation konzentriert. Ein weiterer Kritikpunkt besteht darin, dass zwar konzeptuell immer wieder betont wird, dass bei *data-based decision making* ein weiter Datenbegriff zugrunde gelegt wird und explizit z.B. auch qualitative Daten adressiert werden (z.B. Schildkamp, 2019). Gleichzeitig steht aber die Nutzung quantitativer Leistungsdaten aus standardisierten Verfahren, die mit dem Ziel verbesserter Leistungen von Schüler*innen verknüpft werden, häufig im Mittelpunkt (Mandinach & Schildkamp, 2021a). Der Kritikpunkt dieser Engführung trifft auch auf die vorliegende Arbeit zu, da in den Studien zu Forschungsfrage 1 die Rezeption und Interpretation in Bezug auf Vergleichsarbeiten und technologiebasiertes formatives Assessment untersucht wurde, also jeweils grafisch aufbereitete quantitative Leistungsdaten aus standardisierten Verfahren.

Die Datenrezeption wurde im Rahmen dieser Arbeit konzeptuell eng mit der Komplexität auf der Basis der drei Stufen von *graph literacy* verknüpft und anschließend entsprechend empirisch operationalisiert. Diese Verknüpfung wird zwar auch in vielen anderen Studien vorgenommen (vgl. Kapitel 3.3), entspricht aber einer normativen Setzung, bei der allgemeine Ansätze von *graph literacy* (z.B. Friel et al., 2001) mehr oder weniger direkt auf die

Datenrezeption von Lehrpersonen bezogen werden. Dementsprechend kann zum einen hinterfragt werden, ob die Datenrezeption von Lehrkräften nur dann als adäquat bezeichnet werden kann, wenn die höchste Stufe von *graph literacy* erreicht wird. Zum anderen erlaubt dieser Fokus auch keine Aussagen über die Konsistenz zwischen Rezeption, Interpretation und konstruierten Anschlusshandlungen. Um dies an einem Beispiel zu verdeutlichen: Angenommen, eine Lehrpersonen rezipiert in Aufgabenergebnissen von technologiebasiertem formativen Assessment, wer am schlechtesten in der Lerngruppe abgeschnitten hat (Relationierung von Datenpunkten, mittlere Stufe von *graph literacy*), macht als mögliche Ursache mangelndes aber notwendiges Vorwissen aufgrund einer längeren Krankheitsphase als Muster aus (Interpretation der Information mit Rückgriff auf Kontextwissen) und bietet daraufhin der Schülerin entsprechende Materialien an (Anschlusshandlung). Dann hat die Lehrperson zwar bei der Datenrezeption die höchste Stufe von *graph literacy* nicht erreicht, scheint aber dennoch eine wichtige Information aus den Daten generiert, diese konsistent und plausibel interpretiert und in eine schlüssige Anschlusshandlung transformiert zu haben.

Als eine Stärke dieser Arbeit in Bezug auf Forschungsfrage 1 wurde der Fokus auf ökologische Validität, also der Fokus auf die alltägliche Rezeption und Interpretation von Lehrpersonen, und der Zugang über die Methode des lauten Denkens herausgestellt. Dabei konnte jedoch aus forschungspraktischen Gründen kein direktes *experience sampling* im Sinne der echten, alltäglichen, direkten Erfassung von Verhalten oder Empfinden von Personen in Echtzeit bzw. enger zeitlicher Nähe (Myin-Germeys & Kuppens, 2022) umgesetzt werden. Dies könnte in zukünftigen Studien stärker berücksichtigt werden.

Im Rahmen dieser Arbeit wurde immer wieder die Unterscheidung zwischen der Kompetenz einer Person erfasst in einem Test und der Performanz im Sinne des tatsächlichen alltäglichen Verhaltens getroffen. Diese Unterscheidung kann aus der Sicht verschiedener kompetenztheoretischer Ansätze diskutiert werden: Einige Ansätze fassen Kompetenz als Fähigkeit im Sinne einer Disposition und Performanz im Sinne der Realisation dieser Disposition im konkreten Verhalten auf (Klieme et al., 2008; Weinert, 2001). Andere Ansätze verknüpfen beide Konzepte und definieren Kompetenz als erfolgreiche Performanz in der Praxis (Blömeke et al., 2015). Gleichzeitig gibt es auch Positionierungen, die weniger dichotom sind und Kompetenz als Kontinuum zwischen kognitiven und affektiv-motivationalen Dispositionen, die vermittelt über die Wahrnehmung, Interpretation und Entscheidungen in konkreten Situationen in der Performanz einer Person resultieren, konzeptualisieren (Blömeke et al., 2015). Die Unterscheidung zwischen Kompetenz und Performanz bzw. die entsprechende Begriffsverwendung im Rahmen dieser Arbeit wurde vorgenommen, um zu

formulieren, dass von der Kompetenz einer Person, erfasst in einem *data literacy*-Test, nicht unbedingt direkt auf die Performanz im Sinne des alltäglichen Verhaltens und Denkens geschlossen werden kann. Um konkretere Einblicke in alltagsnahe Kognitionen zu gewinnen (und in Artikel 1 die Zusammenhänge zwischen Testscores und verbalisierten Kognitionen zu explorieren), wurde lautes Denken als methodischer Schwerpunkt für die Arbeit gewählt.

Eine weitere zentrale Limitation dieser Arbeit betrifft den Fokus der Digitalisierung und Datafizierung und die Ausblendung aktueller Innovationen in diesem Bereich: Der Bereich maschinellen Lernens und künstlicher Intelligenz und deren Implikationen für die Kompetenzen von Lehrpersonen wurde nur marginal berücksichtigt, weshalb man mangelnde Aktualität mit Blick auf den prinzipiellen technologischen Fortschritt kritisieren kann. Letzteres betrifft z.B. den Bereich *learning analytics* mit Blick auf Lehrkräfte und deren Implikationen für die Professionalität und Professionalisierung von Lehrkräften (Molenaar, 2022; Selwyn, 2019). Mit Blick auf die aktuelle Praxis in Schulen in Deutschland scheint dieser Schwerpunkt, verbunden mit dem Fokus der ökologischen Validität, allerdings rechtfertigbar. Denn technologiebasiertes formatives Assessment kann durchaus als aktuelle Innovation gesehen werden und es erscheint nicht absehbar, welche Anwendungen von *learning analytics*, maschinellem Lernen und künstlicher Intelligenz allgemein in Zukunft wie implementiert werden (z.B. Jude et al., 2020, 2023).

9.3. Anschließende Forschungsfragen

Unmittelbar an die erste (explorative) Forschungsfrage dieser Arbeit schließt sich die Frage nach der Konsistenz zwischen Rezeption, Interpretation und der Konstruktion pädagogisch-didaktischer Anschlusshandlungen an. Hier könnten (die vorliegenden) Daten des lauten Denkens z.B. dahingehend untersucht werden, inwiefern formulierte Handlungsmaßnahmen an vorher Verbalisiertes z.B. im Rahmen der Datenrezeption anknüpfen. Allerdings dürfte auf der Basis der methodologischen Annahmen von lautem Denken hierbei von nicht Verbalisiertem nicht auf das Nicht-Vorhandensein in den Kognitionen geschlossen werden (Fox et al., 2011).

Eine weitere unmittelbar anschließende Forschungsfrage an diese Arbeit betrifft den Zusammenhang zwischen der allgemeinen *data literacy* von Lehrkräften und ihrer Performanz im Umgang mit Daten im Alltag sowie möglicher Moderatoren: Beispielsweise könnte untersucht werden, inwiefern sich die Förderung der generischen *data literacy* auf den konkreten Umgang mit bestimmten “Verfahren” wie Vergleichsarbeiten oder Datenarten wie Feedback zur Unterrichtsqualität auswirkt. Als mögliche Moderatoren liegen im Anschluss an die bisherige Forschung etwa motivationale Überzeugungen und Einstellungen gegenüber

datengestützten Entscheidungen nahe (z.B. Prenger & Schildkamp, 2018). In diesem Kontext wäre auch die Untersuchung des Verhältnisses von Kontextwissen und Domänen professionellen Wissens nach Shulman (1987) und *data literacy* von Relevanz. Die Erkenntnisse und Implikationen dieser skizzierten Forschungsfragen könnten dann wiederum für Professionalisierungsprozesse bei angehenden Lehrpersonen fruchtbar gemacht werden, etwa inwiefern die Förderung von *data literacy* sinnvoll gemeinsam mit fachlichem, fachdidaktischem usw. Wissen angebahnt wird.

Verschiedene Autor*innen werfen, im Anschluss an eigene Studien oder auch aufgrund konzeptueller Überlegungen, immer wieder die Frage auf, inwiefern von einer mangelnden Passung zwischen den Kompetenzen von Lehrkräften im Umgang mit Daten und den zur Verfügung stehenden Daten bzw. deren Aufbereitung und Darstellung in den Rückmeldungen ausgegangen werden muss (z.B. Altrichter et al., 2016; Goffin et al., 2023; Gutwirth et al., 2021; van der Kleij & Eggen, 2013). Folgt man dieser Positionierung, wofür die entsprechenden Studien sprechen, liegt es zunächst nahe, an den Kompetenzen der Lehrkräfte und der entsprechenden Förderung anzusetzen, wie es z.B. auch diese Arbeit tut (vgl. Forschungsfrage 2). Ergänzend dazu scheint es gleichzeitig vielversprechend, die Angebotsseite zu fokussieren und zu untersuchen, welche Daten und insbesondere welche Aufbereitungen und Darstellungen von Daten für Lehrkräfte hilfreich und intuitiv verständlich sind. Dabei scheint es sinnvoll, die Erkenntnisse zur Perspektive von Lehrkräften, die etwa im Rahmen von Interviewstudien und lautem Denken gewonnen wurden, zu berücksichtigen. Anknüpfend an die vorliegende Arbeit betrifft dies z.B. die Frage, wie Aufgabenergebnisse einer Klasse so dargestellt werden können, dass Lehrkräfte etwa im synchronen Unterricht in Dashboards oder auch für die Unterrichtsplanung im Kontext von formativem Assessment vergleichsweise niedrigschwellig rezipieren und interpretieren können, wer welche Lernziele bereits erreicht hat und wie z.B. sinnvoll Lerngruppen gebildet werden könnten (vgl. Fehleranalyse der Lehrkräfte in Artikel 2). Weiterhin könnte hier experimentell untersucht werden, inwiefern unterschiedliche Visualisierungen die Adressierung einer bestimmten Bezugsnorm bei der Interpretation beeinflussen (vgl. Artikel 1). Die Frage, *welche* Bezugsnorm(en) hierbei in welchem Kontext visuell in den Vorder- bzw. Hintergrund gerückt werden sollte(n), müsste dabei auf einer normativen Ebene diskutiert werden, da alle Bezugsnormen durch die unterschiedlichen Relationierungen notwendigerweise blinde Flecken für die Beurteilung aufweisen (Rheinberg, 2014). Bei der Datenaufbereitung, -triangulation und Gestaltung der Rückmeldungen stellt sich auch die Frage, z.B. im Kontext von *learning analytics*, inwiefern Technologie Lehrkräfte unterstützen könnte (Dimitriadis et al., 2021) und wie eine gelingende Interaktion von Mensch und Technologie aussehen könnte (Mandinach & Schildkamp, 2021; Molenaar, 2022). Aufgrund der zentralen

Rolle, die Lehrpersonen im Prozess einnehmen, scheint es notwendig, ihre Perspektive frühzeitig und stark einzubinden, z.B. durch die Ko-Konstruktion von Technologie zwischen Entwickler*innen, Forschenden und Lehrkräften (Penuel et al., 2007) oder *design-based research* (The Design-Based Research Collective, 2003).

Im Kontext von Digitalisierung und Technologie scheint auch auf einer konzeptuellen Ebene weiterer Forschungsbedarf zu liegen: Wenn auf der Basis des weiten Datenbegriffs *big data* in datengestützte Entscheidungen für Lehrkräfte eingeschlossen wird, also z.B. die Nutzung von Logdaten aus Lernmanagementsystemen, was aktuelle Ansätze zwar tun, aber nicht weiter ausführen (z.B. Mandinach & Gummer, 2016; Schildkamp, 2019), stellt sich die Frage, welche Implikationen sich für das Konstrukt *data literacy* daraus ergeben.

9.4. Implikationen für die schulische Praxis sowie die Aus- und Weiterbildung von Lehrpersonen

Die Ergebnisse dieser Arbeit deuten, wie andere Studien auch, darauf hin, dass Lehrkräfte insgesamt eher Schwierigkeiten haben, Schule und Unterricht datenbasiert zu gestalten und weiterzuentwickeln und die dafür notwendigen Kompetenzen nicht unbedingt vorausgesetzt werden können, letztere aber bei (angehenden) Lehrkräften nachweislich angebahnt und gefördert werden können. Wenn datenbasierte Entscheidungen als Teil professionellen Handelns von Lehrkräften erwartet werden, stellt sich daher die Frage nach Unterstützungssystemen für Lehrpersonen in der Praxis. Ein aus der Wissenschaft entwickeltes und empirisch in seiner Wirksamkeit überprüftes Konzept zur Unterstützung von Lehrpersonen sind sogenannte *data teams* (Schildkamp et al., 2016, 2018, 2019): *Data teams* sind eine Form professioneller Lerngemeinschaften und bestehen aus mehreren Lehrkräften einer Schule, Personen aus dem Schulleitungsteam (oder die Schulleitung selbst) sowie einer externen Person mit Expertise im Umgang mit Daten, in der Regel ein*e Wissenschaftler*in. Dieses Team trifft sich monatlich über ein bis zwei Schuljahre hinweg und bearbeitet auf der Basis eines strukturierten Ansatzes gemeinsam ein schulinternes Problem, für das sich das Team selbst entscheidet und zu dem das Team in einen datengestützten Schul- bzw. Unterrichtsentwicklungsprozess eintreten möchte. Die externe Person übernimmt hierbei eine Moderationsrolle und bringt zudem ihre Expertise im Umgang mit Daten ein. Abgesehen von dem rein zeitlichen und organisatorischen Aufwand für alle Beteiligten gelten hier u.a. die Fähigkeit zur Kooperation, ein gemeinsames Ziel sowie entsprechende Einstellungen als relevante Faktoren (Schildkamp & Poortman, 2015). Durchgeführte Begleitstudien in den Niederlanden zeigen nicht für alle Schulen signifikante positive Effekte, aber in etwas mehr als der Hälfte der Schulen konnten signifikante positive Effekte auf die Leistungen der Schüler*innen nachgewiesen werden (Poortman & Schildkamp, 2016).

Wenn datenbasierte Entscheidungen als Teil professionellen Handelns von Lehrkräften erwartet werden, scheint eine stärkere Anbahnung von *data literacy* über die verschiedenen Phasen der Lehrer*innenbildung hinweg notwendig, um diese systemische Innovation hinsichtlich der Entwicklung der entsprechenden professionellen Kompetenzen zu begleiten. In den Standards für die Lehrerbildung im Bereich Bildungswissenschaften (KMK, 2004, in der Fassung 2022) finden sich dazu Anklänge, wie z.B. der Einsatz von Lernprozessdiagnostik oder die Nutzung von Daten aus Evaluationen. Ergänzend dazu liegen erste konzeptuelle Arbeiten (z.B. Beck et al., 2019) sowie empirisch überprüfte Interventionen dazu vor (z.B. Merk et al., 2020; Reeves & Honig, 2015). Bei der Adressierung von *data literacy* im Lehramtsstudium scheint es zudem sinnvoll, *data literacy* mit fachlichen, fachdidaktischen und bildungswissenschaftlichen Studienanteilen zu verbinden und aufgrund konzeptueller Ähnlichkeiten zu versuchen, Synergien mit der Anbahnung der diagnostischen Kompetenz von Lehrpersonen zu induzieren. Denn damit sich die Potenziale einer datengestützten Gestaltung und Entwicklung von Schule und Unterricht entfalten können und mögliche dysfunktionale Wirkungen minimiert werden, sind hinreichend ausgeprägte Kompetenzen von Lehrpersonen und Unterstützungssysteme in der schulischen Praxis notwendige Voraussetzungen. Wirksame Interventionen zur Förderung von *data literacy* und evaluierte Konzepte für Unterstützungssysteme, z.B. die genannten *data teams*, scheinen hierfür geeignete Anknüpfungspunkte aus der Wissenschaft sowohl für die schulische Praxis als auch andere wichtige Akteure wie etwa die Landesinstitute und die Bildungspolitik.

Literatur

Altrichter, H., & Maag Merki, K. (Hrsg.). (2016). *Handbuch Neue Steuerung im Schulsystem* (2. Aufl.). Springer VS. https://doi.org/10.1007/978-3-531-18942-0

Altrichter, H., Moosbrugger, R., & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2. Aufl., S. 235–277). Springer VS. https://doi.org/10.1007/978-3-531-18942-0_9

Ansyari, M. F., Groot, W., & De Witte, K. (2020). Tracking the process of data use professional development interventions for instructional improvement: A systematic literature review. *Educational Research Review*, *31*, 100362. https://doi.org/10.1016/j.edurev.2020.100362

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press.

Balfanz, R., & Byrnes, V. (2018). Using Data and the Human Touch: Evaluating the NYC Inter-Agency Campaign to Reduce Chronic Absenteeism. Journal of Education for Students Placed at Risk (JESPAR), 23(1–2), 107–121. https://doi.org/10.1080/10824669.2018.1435283

Beck, J. S., Morgan, J. J., Whitesides, H., Riddle, D., & Brown, N. (2019, April). Differentiating between data literacy and assessment literacy: A systematic review of research. *Paper presented at the annual meeting of the American Educational Research Association*. Abstract retrieved from http://tinyurl.com/yb8r3fmb.

Beck, J. S., & Nunnaley, D. (2020). A continuum of data literacy for teaching. *Studies in Educational Evaluation*, 100871. https://doi.org/10.1016/j.stueduc.2020.100871

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

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, *52*(5), 861–893. https://doi.org/10.3102/0002831215599251

Bez, S., Burkart, F., Tomasik, M. J., & Merk, S. (revise and resubmit). How do teachers make sense of technology-based formative assessments in their daily practice? Results from process mining of think-aloud data. *Learning and Instruction*.

Bez, S., Poindl, S., Bohl, T., & Merk, S. (2021). Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien. *Zeitschrift für Pädagogik*, *67*(4), 551–572.

Bez, S., Tomasik, M. J., & Merk, S. (2023). Data-based decision making in einer digitalen Welt: Data Literacy von Lehrpersonen als notwendige Voraussetzung. In K. Scheiter & I. Gogolin (Hrsg.), *Bildung für eine digitale Zukunft* (S. 339–362). Springer VS. https://doi.org/10.1007/978-3-658-37895-0_14

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., Smith, K., Timperley, H., Volante, L., & Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education*, *13*(1), 117–140. https://doi.org/10.1177/1478210314566733

Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, *28*(2), 217–236. https://doi.org/10.1080/1475939X.2019.1572534

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Brown, C., & Malin, J. R. (2022). *The Emerald handbook of evidence-informed practice in education learning from international contexts*. Emerald Publishing. doi: 10.1108/9781800431416

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Castillo, J. M., March, A. L., Tan, S. Y., Stockslager, K. M., & Brundage, A. (2016). Relationships Between Ongoing Professional Development and Educators' Beliefs Relative to Response to Intervention. *Journal of Applied School Psychology*, *32*(4), 287–312. https://doi.org/10.1080/15377903.2016.1207736

Chick, H., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*, *15*(2), 5–26.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives*, *9*(4), 173–206.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Research in Learning Technology*, *13*(1). https://doi.org/10.3402/rlt.v13i1.10970

Corno, L. (2008). On Teaching Adaptively. *Educational Psychologist*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Cronbach, L. J. (1964). Evaluation for course improvement. In R. W. Heath (Ed.), *New Curricula* (pp. 231–248). Harper & Row.

Cui, Y., & Zhang, H. (2022). Integrating teacher data literacy with TPACK: A self-report study based on a novel framework for teachers' professional development. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.966575

Curcio, F. R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education*, *18*(5), 382–393. https://doi.org/10.2307/749086

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

Datnow, A., & Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *Journal of Educational Change*, *19*(2), 131–152. https://doi.org/10.1007/s10833-018-9323-6

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaft*, *39*(1), 63–83.

Dedering, K., & Kallenbach, L. (2023). Forschungs- und Evidenzbasierung in Schulen. Das Forschungsfeld im Überblick. In K.-S. Besa, D. Demski, J. Gesang, & J.-H. Hinzke (Hrsg.), *Evidenz- und Forschungsorientierung in Lehrer*innenbildung, Schule, Bildungspolitik und -administration: Neue Befunde zu alten Problemen* (S. 125–152). Springer Fachmedien. https://doi.org/10.1007/978-3-658-38377-0_7

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dimitriadis, Y., Martínez-Maldonado, R., & Wiley, K. (2021). Human-Centered Design Principles for Actionable Learning Analytics. In T. Tsiatsos, S. Demetriadis, A. Mikropoulos, & V. Dagdilelis (Hrsg.), *Research on E-Learning and ICT in Education: Technological, Pedagogical and Instructional Perspectives* (S. 277–296). Springer International Publishing. https://doi.org/10.1007/978-3-030-64363-8_15

Dodman, S. L., DeMulder, E. K., View, J. L., Swalwell, K., Stribling, S., Ra, S., & Dallman, L. (2019). Equity Audits as a Tool of Critical Data-Driven Decision Making: Preparing

Teachers to See Beyond Achievement Gaps and Bubbles. Action in Teacher Education, 41(1), 4–22. https://doi.org/10.1080/01626620.2018.1536900

Dodman, S. L., Swalwell, K., DeMulder, E. K., View, J. L., & Stribling, S. M. (2021). Critical data-driven decision making: A conceptual model of data use for equity. *Teaching and Teacher Education*, *99*, 103272. https://doi.org/10.1016/j.tate.2020.103272

Döring, N., Bortz, J., & Pöschl-Günther, S. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. vollständig überarbeitete, aktualisierte und erweiterte Auflage). Springer.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Hrsg.), *Handbook of research on teaching* (S. 392-431). Macmillan.

Du, X., Yang, J., Shelton, B. E., Hung, J.-L., & Zhang, M. (2021). A systematic meta-Review and analysis of learning analytics research. *Behaviour & Information Technology*, *40*(1), 49–62. https://doi.org/10.1080/0144929X.2019.1669712

Dumont, H. (2019). Neuer Schlauch für alten Wein? Eine konzeptuelle Betrachtung von individueller Förderung im Unterricht. *Zeitschrift für Erziehungswissenschaft*, *22*(2), 249–277. https://doi.org/10.1007/s11618-018-0840-0

Dunn, K. E., Airola, D. T., Lo, W.-J., & Garrison, M. (2013). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*(1), 87–98. https://doi.org/10.1016/j.cedpsych.2012.11.002

Dunn, K., & Mulvenon, S. (2009). A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education. *Practical Assessment, Research & Evaluation*, *14*(7). https://doi.org/10.4324/9780203462041_chapter_1

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2017). The effects of a data use intervention on educators' satisfaction and data literacy. *Educational Assessment, Evaluation and Accountability*, *29*(1), 83–105.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J., & Midgley, C. (1983). Expectancies, values and academic behaviors. In J. T. Spense (Ed.), *Achievement and Achievement Motives* (S. 75–146). W. H. Freeman.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Ehren, M. C. M., & Swanborn, M. S. L. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement*, *23*(2), 257–280. https://doi.org/10.1080/09243453.2011.652127

Eickelmann, B. (2018). Digitalisierung in der schulischen Bildung. Entwicklungen, Befunde und Perspektiven für die Schulentwicklung und die Bildungsforschung. In N. McElvany,

F. Schwabe, W. Bos, & H. G. Holtappels (Hrsg.), *Digitalisierung in der schulischen Bildung. Chancen und Herausforderungen* (Bd. 2, S. 11–26). Waxmann.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, *32*(1), 8–21.

Eyal, L. (2012). Digital assessment literacy — the core role of the teacher in a digital environment. *Journal of Educational Technology & Society*, *15*(2), 37–49.

Faber, J. M., Feskens, R., & Visscher, A. J. (2023). A best-evidence meta-analysis of the effects of digital monitoring tools for teachers on student achievement. *School Effectiveness and School Improvement*, *34*(2), 169–188. https://doi.org/10.1080/09243453.2022.2142247

Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education*, *106*, 83–96. https://doi.org/10.1016/j.compedu.2016.12.001

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

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-Based Decision Making in Reading Interventions: A Synthesis and Meta-Analysis of the Effects for Struggling Readers. *The Journal of Special Education*, *52*(3), 174–187. https://doi.org/10.1177/0022466918790001

Filderman, M. J., Toste, J. R., Didion, L., & Peng, P. (2022). Data Literacy Training for K–12 Teachers: A Meta-Analysis of the Effects on Teacher Outcomes. *Remedial and Special Education*, *43*(5), 328–343. https://doi.org/10.1177/07419325211054208

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–158.

Gadermann, A., Guhn, M., & Zumbo, B. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, *17*(1), 1–13.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*(3), 444–457.

Gelderblom, G., Schildkamp, K., Pieters, J., & Ehren, M. (2016). Data-based decision making for instructional improvement in primary education. *International Journal of Educational Research*, *80*, 1–14. https://doi.org/10.1016/j.ijer.2016.07.004

Goffin, E., Janssen, R., & Vanhoof, J. (2022). Teachers' and school leaders' sensemaking of formal achievement data: A conceptual review. *Review of Education*, *10*(1), e3334. https://doi.org/10.1002/rev3.3334

Goffin, E., Janssen, R., & Vanhoof, J. (2023). Principals' and Teachers' Comprehension of School Performance Feedback Reports. Exploring Misconceptions from a User Validity Perspective. *Pedagogische Studiën*, *100*(1), Article 1. https://doi.org/10.59302/ps.v100i1.13991

Gottheiner, D. M., & Siegel, M. A. (2012). Experienced Middle School Science Teachers' Assessment Literacy: Investigating Knowledge of Students' Conceptions in Genetics and Ways to Shape Instruction. *Journal of Science Teacher Education*, *23*(5), 531–557. https://doi.org/10.1007/s10972-012-9278-z

Gottlebe, K., Dietrich, S., Berger, I., Angersbach, C., & Latzko, B. (2023). Diagnostische Praxis digital gestalten – digitale Kompetenzen von Lehrpersonen für die Gestaltung eines lernwirksamen Unterrichts. In S. Ganguin, H. Tiemann, C. W. Glück, & A. Förster (Hrsg.), *Digitalisierung in der Lehrer:innenbildung: Praxis digital gestalten* (S. 65–85). Springer Fachmedien. https://doi.org/10.1007/978-3-658-41637-9_4

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Groß Ophoff, J. (2013a). Der Effekt der Bezugsnormorientierung auf die Reflexion und Nutzung von Rückmeldungen aus Vergleichsarbeiten. *Empirische Pädagogik*, *27*(4), 442–458.

Groß Ophoff, J. (2013b). *Lernstandserhebungen: Reflexion und Nutzung*. Münster u.a.: Waxmann.

Groß Ophoff, J., & Cramer, C. (2022). The Engagement of Teachers and School Leaders with Data, Evidence and Research in Germany. In C. Brown & J. R. Malin (Hrsg.), *The Emerald Handbook of Evidence-Informed Practice in Education* (S. 175–195). Emerald Publishing Limited. https://doi.org/10.1108/978-1-80043-141-620221026

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*(8), 1526–1553. https://doi.org/10.1080/00949655.2019.1590574

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *The British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261.

Gutwirth, G., Goffin, E., & Vanhoof, J. (2021). Sensemaking unraveled: How teachers process school performance feedback data. *Studia Paedagogica*, *26*(4), Article 4. https://doi.org/10.5817/SP2021-4-4

Hall, T. E., Cohen, N., Vue, G., & Ganley, P. (2015). Addressing Learning Disabilities With UDL and Technology: Strategic Reader. *Learning Disability Quarterly*, *38*(2), 72–83. https://doi.org/10.1177/0731948714544375

Hamilton, L., Halversion, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009a). *Using student achievement data to support instructional decision making*. Washington D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Hamilton, L., Stecher, B. M., & Yuan, K. (2009b). *Standards-Based reform in the United States: History, research, and future directions*. Center on Education Policy. https://www.rand.org/pubs/reprints/RP1384.html.

Hamilton, V. M., & Reeves, T. D. (2021). Relationships between Course Taking and Teacher Self-Efficacy and Anxiety for Data-Driven Decision Making. *The Teacher Educator*, *0*(0), 1–19. https://doi.org/10.1080/08878730.2021.1965682

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*(2). https://doi.org/10.25656/01:18004

Hartmann, C., Rummel, N., & Bannert, M. (2022). Using HeuristicsMiner to Analyze Problem-Solving Processes: Exemplary Use Case of a Productive-Failure Study. *Journal of Learning Analytics*, *9*(2), Article 2. https://doi.org/10.18608/jla.2022.7363

Hartong, S. (2019). Bildung 4.0? Kritische Überlegungen zur Digitalisierung von Bildung als erziehungswissenschaftliches Forschungsfeld. *Zeitschrift für Pädagogik*, *65*(3), 424–444.

Hartong, S., Breiter, A., Jarke, J., & Förschler, A. (2019). Digitalisierung von Schule, Schulverwaltung und Schulaufsicht. In T. Klenk, F. Nullmeier, & G. Wewer (Hrsg.), *Handbuch Digitalisierung in Staat und Verwaltung* (S. 1–10). Springer Fachmedien. https://doi.org/10.1007/978-3-658-23669-4_43-1

Hasselhorn, M., Decristan, J., & Klieme, E. (2019). Individuelle Förderung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, H. Solga, et al. (Hrsg.), *Das Bildungswesen in Deutschland: Bestand und Potenziale* (S. 375–401). Klinkhardt.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Hebbecker, K., Förster, N., Forthmann, B., & Souvignier, E. (2022). Data-based decision-making in schools: Examining the process and effects of teacher support. *Journal of Educational Psychology*, *114*(7), 1695–1721. https://doi.org/10.1037/edu0000530

Heitink, M. C., van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, *17*, 50–62. https://doi.org/10.1016/j.edurev.2015.12.002

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., et al. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research*, *7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Hellrung, K., & Hartig, J. (2013). Understanding and using feedback – A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190.

Helmke, A., & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Beat, N. Landwehr & W. Böttcher (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127–151). hep.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, *76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hoogland, I., Schildkamp, K., van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., & Dijkstra, A. M. (2016). Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teaching and Teacher Education*, *60*, 377–386. https://doi.org/10.1016/j.tate.2016.07.012

Hosenfeld, I., & Groß Ophoff, J. (2007). Nutzung und Nutzen von Evaluationsstudien in Schule und Unterricht. *Empirische Pädagogik*, *27*(4), 352–367.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Huguet, A., Marsh, J. A., & Farrell, C. (2014). Building Teachers' Data-use Capacity: Insights from Strong and Developing Coaches. *Education Policy Analysis Archives*, *22*(0), 52. https://doi.org/10.14507/epaa.v22n52.2014

Institut für Bildungsanalysen Baden-Württemberg (Hrsg.) (2019). *Vergleichsarbeiten VERA 3. Nutzung der Ergebnisse im Rahmen der Qualitätssicherung in Schulen*. https://ibbw.kultus-bw.de/site/pbs-bw-km-root/get/documents_E-587667235/KULTUS.Dachmandant/KULTUS/Dienststellen/ls-bw/Lernstandserhebungen/dokumente/vera3docs/IBBW/v3_Handreichung_Nutzung_Ergebnisse.pdf

Jarke, J., & Macgilchrist, F. (2021). Dashboard stories: How narratives told by predictive analytics reconfigure roles, risk and sociality in education. *Big Data & Society*, *8*(1), 20539517211025560. https://doi.org/10.1177/20539517211025561

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Hrsg.), *Data Driven Approaches in Digital Education* (S. 82–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-66610-5_7

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate: preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (S. 31–40). https://doi.org/10.1145/3170358.3170421

Jude, N., Ziehm, J., Goldhammer, F., Drachsler, H., & Hasselhorn, M. (2020). *Digitalisierung an Schulen: eine Bestandsaufnahme*. Frankfurt a. M.: DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation. https://www.pedocs.de/volltexte/2020/20522/pdf/Jude_et_al_2020_Digitalisierung_an_Schulen.pdf.

Jude, N., Ziehm-Eicher, J., Goldhammer, F., Drachsler, H., & Hasselhorn, M. (2023). Digitalisierung und Diagnostik in Schulen – Herausforderungen für Bildungspraxis und Bildungsforschung. In K. Scheiter & I. Gogolin (Hrsg.), *Bildung für eine digitale Zukunft* (S. 275–292). Springer Fachmedien. https://doi.org/10.1007/978-3-658-37895-0_11

Jungjohann, J., Gebhardt, M., & Scheer, D. (2022). Understanding and improving teachers' graph literacy for data-based decision-making via video intervention. *Frontiers in Education*, *7*. https://www.frontiersin.org/articles/10.3389/feduc.2022.919152

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kärner, T., Keller, T., Schneider, A., Albaner, D., & Schumann, S. (2021). Ein Rahmenmodell zur Gestaltung technologisch unterstützter adaptiver Lehr- und Lernprozesse. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, *117*(3), 351–371. https://doi.org/10.25162/zbw-2021-0016

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data. An introduction to cluster analysis* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316801

Kennedy-Clark, S., & Reimann, P. (2022). Knowledge types in initial teacher education: A multi-dimensional approach to developing data literacy and data fluency. *Learning: Research and Practice*, *8*(1), 42–58. https://doi.org/10.1080/23735082.2021.1957140

Kettenring, J. R. (2006). The Practice of Cluster Analysis. *Journal of Classification*, *23*(1), 3–30. https://doi.org/10.1007/s00357-006-0002-6

Keuning, T., & van Geel, M. (2021). Differentiated teaching with adaptive learning systems and teacher dashboards: The teacher still matters most. *IEEE Transactions on Learning Technologies*, *14*(2), 201–210. https://doi.org/10.1109/TLT.2021.3072143

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

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

Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation*, *56*, 21–31. https://doi.org/10.1016/j.stueduc.2017.11.001

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Hrsg.), *Assessment of Competencies in Educational Contexts* (S. 3–22). Hogrefe & Huber.

KMK (2000). *Aufgaben von Lehrerinnen und Lehrern heute—Fachleute für das Lernen*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2000/2000_10_05-Bremer-Erkl-Lehrerbildung.pdf

KMK (2004). *Standards für die Lehrerbildung: Bildungwissenschaften*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

KMK (2016a). *Bildung in der digitalen Welt. Strategie der Kultusministerkonferenz*. https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf.

KMK (2016b). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf

KMK (2018). *Vereinbarung zur Weiterentwicklung von Vergleichsarbeiten (VERA)*. *Beschluss der Kultusministerkonferenz vom 08.03.2012 i. d. F. vom 15.03.2018)*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Knoop-van Campen, C. A. N., & Molenaar, I. (2020). How teachers integrate dashboards into their feedback practices. *Frontline Learning Research*, *8*(4), 37–51. https://doi.org/10.14786/flr.v8i4.641

Knoop-van Campen, C. A. N., Wise, A., & Molenaar, I. (2021). The equalizing effect of teacher dashboards on feedback in K-12 classrooms. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2021.1931346

Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten? Datenkompetenz von Lehrkräften und die Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten*. Waxmann.

Koch, U. (2013). Datenauswertungskompetenzen von Lehrkräften. In J. U. Hense, W. Böttcher, S. Rädiker & T. Widmer (Hrsg.), *Forschung über Evaluation: Bedingungen, Prozesse, Wirkungen* (S. 21–41). Waxmann.

Krein, U., & Schiefner-Rohs, M. (2021). Data in Schools: (Changing) Practices and Blind Spots at a Glance. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.672666

Lai, M. K., & Schildkamp, K. (2013). Data-based decision making in Education: An Overview. In K. Schildkamp, M. K. Lai, & L. Earl (Hrsg.), *Data-based decision making in education. Challenges and opportunities* (S. 9–21). Springer.

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review. *Applied Measurement in Education*, *33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Lembke, E. S., McMaster, K. L., Smith, R. A., Allen, A., Brandes, D., & Wagner, K. (2018). Professional Development for Data-Based Instruction in Early Writing: Tools, Learning, and Collaborative Support. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *41*(2), 106–120. https://doi.org/10.1177/0888406417730112

Maag Merki, K. (2016). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 151–181). Springer Fachmedien.

https://doi.org/10.1007/978-3-531-18942-0_6

Maier, U., & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen. *Die Deutsche Schule*, *104*(1), 88–99.

Maier, U., Metz, K., Bohl, T., Kleinknecht, M., & Schymala, M. (2012). Vergleichsarbeiten als Instrument der datenbasierten Schul- und Unterrichtsentwicklung in Gymnasien. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 197–224). Springer VS.

Mandinach, E. B., & Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educational Researcher*, *42*(1), 30–37. https://doi.org/10.3102/0013189X12459803

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376.

Mandinach, E. B., & Schildkamp, K. (2021a). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation*, *69*, 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Mandinach, E. B., & Schildkamp, K. (2021b). The complexity of data-based decision making: An introduction to the special issue. *Studies in Educational Evaluation*, *69*, 100906. https://doi.org/10.1016/j.stueduc.2020.100906

Mannhardt, F., & Janssenswillen, G. (2023). *heuristicsmineR* (0.3.0) [Computer software].

Marsh, H. W. (2007). Application of Confirmatory Factor Analysis and Structural Equation Modeling in Sport and Exercise Psychology. In G. Tenenbaum & R. C. Eklund (Hrsg.), *Handbook of Sport Psychology* (S. 774–798). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118270011.ch35

Marsh, J. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*, 1–48.

McArdle, J. J., & Grimm, K. J. (2010). Five Steps in Latent Curve and Latent Change Score Modeling with Longitudinal Data. In K. van Montfort, J. H. L. Oud, & A. Satorra (Hrsg.), *Longitudinal Research with Latent Variables* (S. 245–273). Springer. https://doi.org/10.1007/978-3-642-11760-2_8

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

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

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. Office of Planning, Evaluation and Policy

Development, US Department of Education. https://www2.ed.gov/rschstat/eval/data-to-inform-instruction/report.pdf.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Merk, S., Poindl, S., Wurster, S., & Bohl, T. (2020). Fostering aspects of pre-service teachers' data literacy: Results of a randomized controlled trial. *Teaching and Teacher Education*, *91*, 103043.

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

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

Muthén, L. K., & Muthén, B. (2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Myin-Germeys, I., & Kuppens, P. (2022). *The Open handbook of experience sampling methodology. A step-by-step guide to designing, conducting, and analyzing ESM studies* (2. Aufl.). Center for Research on Experience Sampling and Ambulatory Methods Leuven.

Nagengast, B., Marsh, H., Scalas, L. F., Xu, K., Hau, K.-T., & Trautwein, U. (2011). Who Took the "×" out of Expectancy-Value Theory? *Psychological Science*, *22*, 1058–1066. https://doi.org/10.1177/0956797611415540

Oslund, E. L., Elleman, A. M., & Wallace, K. (2021). Factors Related to Data-Based Decision-Making: Examining Experience, Professional Development, and the Mediating Effect of Confidence on Teacher Graph Literacy. *Journal of Learning Disabilities*, *54*(4), 243–255. https://doi.org/10.1177/0022219420972187

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Hrsg.), *Understanding and investigating response processes in validation research* (S. 211–228). Springer.

Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education*, *8*. https://www.frontiersin.org/articles/10.3389/feduc.2023.1217167

Peek, R., & Dobbelstein, P. (2006). Benchmarks als Input für die Schulentwicklung – Das Beispiel der Lernstandserhebungen in Nordrhein-Westfalen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 41–58). Waxmann.

Penuel, W., Roschelle, J., & Shechtman, N. (2007). Designing Formative Assessment Software with Teachers: An Analysis of the Co-Design Process. *Research and Practice*

in Technology Enhanced Learning, *2*, 51–74. https://doi.org/10.1142/S1793206807000300

Pierce, R., & Chick, H. (2011). Teachers' intentions to use national literacy and numeracy assessment data: A pilot study. *The Australian Educational Researcher*, *38*, 433–447. https://doi.org/10.1007/s13384-011-0040-x

Pierce, R., Chick, H., Watson, J., Les, M., & Dalton, M. (2014). A statistical literacy hierarchy for interpreting educational system data: *Australian Journal of Education*, *58*(2), 195–217. https://doi.org/10.1177/0004944114530067

Plante, I., O'Keefe, P. A., & Théorêt, M. (2013). The relation between achievement goal and expectancy-value theories in predicting achievement-related outcomes: A test of four theoretical conceptions. *Motivation and Emotion*, *37*(1), 65–78. https://doi.org/10.1007/s11031-012-9282-9

Poortman, C. L., & Schildkamp, K. (2016). Solving student achievement problems with a data use intervention for teachers. *Teaching and Teacher Education*, *60*, 425–433. https://doi.org/10.1016/j.tate.2016.06.010

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: A psychological perspective on the role of the teacher. *Educational Psychology*, *38*(6), 734–752. https://doi.org/10.1080/01443410.2018.1426834

Ratner, H., Andersen, B. L., & Madsen, S. R. (2019). Configuring the teacher as data user: public-private sector mediations of national test data. *Learning, Media and Technology*, *44*(1), 22–35. https://doi.org/10.1080/17439884.2018.1556218

Reeves, T. D. (2017). Equipping preservice elementary teachers for data use in the classroom. *Action in Teacher Education*, *39*(4), 361–380. https://doi.org/10.1080/01626620.2017.1336131

Reeves, T. D., & Chiang, J.-L. (2017). Building Pre-service Teacher Capacity to Use External Assessment Data: An Intervention Study. *The Teacher Educator*, *52*(2), 155–172. https://doi.org/10.1080/08878730.2016.1273420

Reeves, T. D., & Chiang, J.-L. (2018). Online interventions to promote teacher data-driven decision making: Optimizing design to maximize impact. *Studies in Educational Evaluation*, *59*, 256–269.

Reeves, T. D., & Chiang, J.-L. (2019). Effects of an asynchronous online data literacy intervention on pre-service and in-service educators' beliefs, self-efficacy, and practices. *Computers & Education*, *136*, 13–33. https://doi.org/10.1016/j.compedu.2019.03.004

Reeves, T. D., & Honig, S. L. (2015). A classroom data literacy intervention for pre-service teachers. *Teaching and Teacher Education*, *50*, 90–101.

Reeves, T. D., Onder, Y., & Abdi, B. (2020). Validation of the Data-Driven Decision-Making Efficacy and Anxiety Inventory (3D-MEA) with U.S. Pre-Service Teachers. *Mid-Western Educational Researcher*, *32*(4), 286–303.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 239–257. https://doi.org/10.1007/s11412-009-9070

Rheinberg, F. (2011). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (3. Aufl., S. 59–71). Beltz.

Richter, D., Böhme, K., Becker, M., Pant, H., & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik, 60*, 225–244.

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rollett, W., Bijlsma, H., & Röhl, S. (2021). Student feedback on teaching in schools: Current state of research and future perspectives. In W. Rollett, H. Bijlsma, & S. Röhl (Hrsg.), *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers* (S. 259–271). Springer International Publishing. https://doi.org/10.1007/978-3-030-75150-0_16

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273.

Schildkamp, K., Handelzalts, A., Poortman, C. L., Leusink, H., Meerdink, M., Smit, M., Ebbeler, J., & Hubers, M. D. (2018). *The Data Team™ Procedure: A Systematic Approach to School Improvement* (K. Schildkamp, A. Handelzalts, C. L. Poortman, H. Leusink, M. Meerdink, M. Smit, J. Ebbeler, & M. D. Hubers, Hrsg.). Springer International Publishing. https://doi.org/10.1007/978-3-319-58853-7_9

Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation*, *42*, 15–24. https://doi.org/10.1016/j.stueduc.2013.10.007

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3), 482–496. https://doi.org/10.1016/j.tate.2009.06.007

Schildkamp, K., Lai, M. K., & Earl, L. (2013). *Data-based Decision Making in Education. Challenges and Opportunities*. Springer.

Schildkamp, K., & Poortman, C. (2015). Factors Influencing the Functioning of Data Teams. *Teachers College Record*, *117*(4).

https://www.tcrecord.org/content.asp?contentid=17851

Schildkamp, K., Poortman, C. L., Ebbeler, J., & Pieters, J. M. (2019). How school leaders can build effective data teams: Five building blocks for a new wave of data-informed decision making. *Journal of Educational Change*, *20*(3), 283–325. https://doi.org/10.1007/s10833-019-09345-3

Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2016). Data teams for school improvement. *School Effectiveness and School Improvement*, *27*(2), 228–254. https://doi.org/10.1080/09243453.2015.1056192

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement*, *28*(2), 242–258. https://doi.org/10.1080/09243453.2016.1256901

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Schliesing, A. C. (2017). *Rückmeldungen aus Vergleichsarbeiten (VERA). Eine methodenintegrative Studie zur Gestaltung und Rezeption von VERA-Rückmeldungen*. Dissertation, Humboldt-Universität zu Berlin.

Schneewind, J. (2007). Erfahrungen mit Erlebnisrückmeldungen im Projekt BeLesen – Ergebnisse der Interviewstudie. *Empirische Pädagogik*, *21*(4), 368–382.

Schrader, F.-W., & Praetorius, A.-K. (2018). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost, Jörn R. Sparfeldt, & S. R. Buch (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (5. Aufl., S. 92–98). Beltz.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Hrsg.), *Perspectives on curriculum evaluation* (S. 39–83). Rand McNally.

Selwyn, N. (2019). What's the Problem with Learning Analytics? *Journal of Learning Analytics*, *6*(3), Article 3. https://doi.org/10.18608/jla.2019.63.3

Shulman, L. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (second). SAGE.

Stelter, A., & Miethe, I. (2019). Forschungsmethoden im Lehramtsstudium – aktueller Stand und Konsequenzen. *Erziehungswissenschaft*, *30*(1), 25–33.

Schütze, B., Souvignier, E., & Hasselhorn, M. (2018). Stichwort – Formatives Assessment. *Zeitschrift für Erziehungswissenschaft*, *21*(4), 697–715. https://doi.org/10.1007/s11618-018-0838-7

Schwendimann, B., Rodríguez-Triana, M., Vozniuk, A., Prieto, L., Boroujeni, M. S., Holzer, A., et al. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Siemens, G., & Gašević, D. (2012). Guest editorial - Learning and knowledge analytics. *Educational Technology and Society*, *15*(3), 1–2.

Souvignier, E., Förster, N., & Salaschek, M. (2014). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Lesen und Mathematik. In M. Hasselhorn, W. Schneider, & U. Trautwein (Hrsg.), *Lernverlaufsdiagnostik* (S. 239–256). Hogrefe.

Sonnenberg, C., & Bannert, M. (2019). Using Process Mining to examine the sustainability of instructional support: How stable are the effects of metacognitive prompting on self-regulatory behavior? *Computers in Human Behavior*, *96*, 259–272. https://doi.org/10.1016/j.chb.2018.06.003

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L. (Joy), Mukama, E., Warusavitarana, A., Dona, K. L., Eichhorn, K., Fluck, A., Huang, R., Bridges, S., Lu, J., Ren, Y., Gui, X., Deneen, C. C., Diego, J. S., & Gibson, D. C. (2016). Technology Enhanced Formative Assessment for 21st Century Learning. *Journal of Educational Technology & Society*, *19*(3), 58–71.

Supovitz, J., & Sirinides, P. (2018). The Linking Study: An Experiment to Strengthen Teachers' Engagement with Data on Teaching and Learning. *American Journal of Education*, *124*(2), 161–189. https://doi.org/10.1086/695610

Syring, M., Bohl, T., & Lachner, A. (2022). Digitalisierung in der Schule: Vorschlag eines systematisierenden Rahmenmodells aus schulpädagogischer Perspektive. *Zeitschrift für Bildungsforschung*. https://doi.org/10.1007/s35834-022-00340-y

Tallent-Runnels, M. K., Thomas, J. A., Lan, W. Y., Cooper, S., Ahern, T. C., Shaw, S. M., & Liu, X. (2006). Teaching Courses Online: A Review of the Research. *Review of Educational Research*, *76*(1), 93–135. https://doi.org/10.3102/00346543076001093

Tarkian, J., Maritzen, N., Eckert, M., & Thiel, F. (2019). Vergleichsarbeiten (VERA) – Konzeption und Implementation in den 16 Ländern. In F. Thiel, J. Tarkian, E.-M. Lankes, N. Maritzen, T. Riecke-Baulecke & A. Kroupa (Hrsg.), *Datenbasierte Qualitätssicherung und -entwicklung in Schulen: Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland* (S. 41–103). Springer Fachmedien.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

The Design-Based Research Collective. (2003). Design-Based Research: An Emerging Paradigm for Educational Inquiry. *Educational Researcher*, *32*(1), 5–8.

https://doi.org/10.3102/0013189X032001005

Thiel, F., Tarkian, J., Lankes, E.-M., Maritzen, N., Riecke-Baulecke, T., & Kroupa, A. (Hrsg.). (2019). *Datenbasierte Qualitätssicherung und -entwicklung in Schulen: Eine Bestandsaufnahme in den Ländern der Bundesrepublik Deutschland*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23240-5

Thoren, K., Wißmann, J., Harks, M., Wenger, M., Kinder, A., & Hannover, B. (2020). Förderung von Datennutzungskompetenzen in der Lehramtsausbildung: Konzeption und Evaluation dreier Seminare. In I. Gogolin, B. Hannover, & A. Scheunpflug (Hrsg.), *Evidenzbasierung in der Lehrkräftebildung* (S. 39–71). Springer Fachmedien. https://doi.org/10.1007/978-3-658-22460-8_3

Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, *9*, 1–17. https://doi.org/10.3389/fpsyg.2018.02245

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice*, *32*(1), 46–60.

van der Aalst, W. M. P. (2016). *Process Mining. Data science in action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4

van den Hurk, H. T. G., Houtveen, A. A. M., & van de Grift, W. J. C. M. (2016). Fostering effective teaching behavior through the use of data-feedback. *Teaching and Teacher Education*, *60*, 444–451.

van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation*, *39*(3), 144–152. https://doi.org/10.1016/j.stueduc.2013.04.002

van der Kleij, F. M., Jorine A. Vermeulen, Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

van der Scheer, E. A., & Visscher, A. J. (2016). Effects of an intensive data-based decision making intervention on teacher efficacy. *Teaching and Teacher Education*, *60*, 34–43. https://doi.org/10.1016/j.tate.2016.07.025

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*, *72*(4), 303–308.

van Geel, M., Keuning, T., Visscher, A., & Fox, J.-P. (2017). Changes in educators' data literacy during a data-based decision making intervention. *Teaching and Teacher Education*, *64*, 187–198. https://doi.org/10.1016/j.tate.2017.02.015

van Geel, M., Keuning, T., Visscher, A. J., & Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, *53*(2), 360–394.

van Leeuwen, A. (2015). Learning analytics to support teachers during synchronous CSCL: balancing between overview and overload. *Journal of Learning Analytics*, *2*(2), 138–162. https://doi.org/10.18608/jla.2015.22.11

Vanlommel, K., & Schildkamp, K. (2019). How Do teachers make sense of data in the context of High-Stakes decision making? *American Educational Research Journal*, *56*(3), 792–821. https://doi.org/10.3102/0002831218803891

Vanlommel, K., van Gasse, R., Vanhoof, J., & van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*, 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

Vanlommel, K., van Gasse, R., Vanhoof, J., & van Petegem, P. (2020). Sorting pupils into their next educational track: How strongly do teachers rely on data-based or intuitive processes when they make the transition decision? *Studies in Educational Evaluation*, 100865. https://doi.org/10.1016/j.stueduc.2020.100865

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method. A practical guide to modelling cognitive processes*. London u.a.: Academic Press.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., & De Laet, T. (2020). Learning analytics dashboards: the past, the present and the future. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (S. 35–40). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375462.3375504

Visscher, A. J. (2021). On the value of data-based decision making in education: The evidence from six intervention studies. *Studies in Educational Evaluation*, *69*, 100899. https://doi.org/10.1016/j.stueduc.2020.100899

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. https://doi.org/10.1076/sesi.14.3.321.15842

Wagner, D. L., Hammerschmidt‐Snidarich, S. M., Espin, C. A., Seifert, K., & McMaster, K. L. (2017). Pre-service Teachers' Interpretation of CBM Progress Monitoring Data. *Learning Disabilities Research & Practice*, *32*(1), 22–31. https://doi.org/10.1111/ldrp.12125

Walker, D. A., Reeves, T. D., & Smith, T. J. (2018). Confirmation of the Data-Driven Decision-Making Efficacy and Anxiety Inventory's Score Factor Structure Among

Teachers. *Journal of Psychoeducational Assessment*, *36*(5), 477–491. https://doi.org/10.1177/0734282916682905

Wang, Y. (2021). When artificial intelligence meets educational leaders' data-informed decision-making: A cautionary tale. *Studies in Educational Evaluation*, *69*, 100872. https://doi.org/10.1016/j.stueduc.2020.100872

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Hogrefe & Huber Publishers.

Weijters, A. J. M. M., & Ribeiro, J. T. S. (2011). Flexible Heuristics Miner (FHM). *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 310–317. https://doi.org/10.1109/CIDM.2011.5949453

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction*, *66*, 101303. https://doi.org/10.1016/j.learninstruc.2020.101303

Wurster, S. (2019). Datengestützte Qualitätssicherung und -entwicklung im Schulsystem. In M. Harring, C. Rohlfs, & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S. 765–776). Waxmann.

Wurster, S., Bez, S., & Merk, S. (2023). Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers. *Learning and Instruction*, *88*, 101806. https://doi.org/10.1016/j.learninstruc.2023.101806

Wurster, S., Richter, D., & Lenski, A. E. (2017). Datenbasierte Unterrichtsentwicklung und ihr Zusammenhang zur Schülerleistung. *Zeitschrift für Erziehungswissenschaft*, *20*(4), 628–650. https://doi.org/10.1007/s11618-017-0759-x

Wurster, S., & Richter, D. (2016). Nutzung von Schülerleistungsdaten aus Vergleichsarbeiten und zentralen Abschlussprüfungen für Unterrichtsentwicklung in Brandenburger Fachkonferenzen. *Journal for Educational Research Online*, *8*(3), 159–183.

Wurster, S., Richter, D., & Lenski, A. E. (2017). Datenbasierte Unterrichtsentwicklung und ihr Zusammenhang zur Schülerleistung. *Zeitschrift für Erziehungswissenschaft*, *20*(4), 628–650.

Xuan, Q., Cheung, A., & Sun, D. (2022). The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.990196

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Zhang, Z., & Yuan, K.-H. (2018). Practical statistical power analysis using Webpower and R. ISDSA Press. https://doi.org/10.35566/power

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*(1), 61–70.

Zimmer-Müller, M., Hosenfeld, I., & Koch, U. (2014). Rückmeldungen nach Vergleichsarbeiten in Grund- und Sekundarschulen. In H. Ditton & A. Müller (Hrsg.), *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S. 195–212). Waxmann.

Zlatkin-Troitschanskaia, O. (2016). *Evidence-based actions within the multilevel system of schools—Requirements, processes, and effects (EviS). Special issue editorial*. https://doi.org/10.25656/01:12802

Aufstellung über Anteil und Rolle bei gemeinschaftlichen Publikationen

Artikel 1: Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien

Bez, S., Poindl, S., Bohl, T., & Merk, S. (2021). Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien. *Zeitschrift für Pädagogik*, *67*(4), 551–572. https://doi.org/10.3262/ZP2104551
Status: Publiziert

Autor*in	Position	wissenschaftliche Ideen (in %)	Datenerhebung und -aufbereitung (in %)	Datenanalyse und -interpretation (in %)	Manuskripterstellung und -überarbeitung (in %)
Sarah Bez	1	20	40	80	80
Simone Poindl	2	30	30	0	0
Thorsten Bohl	3	0	0	0	5
Samuel Merk	4	50	30	20	15

Artikel 2: How do teachers make sense of technology-based formative assessments in their daily practice? Results from process mining of think-aloud data

Bez, S., Burkart, F., Tomasik, M. J., & Merk, S. (revise and resubmit). How do teachers make sense of technology-based formative assessments in their daily practice? Results from process mining of think-aloud data. *Learning and Instruction*. [Special issue: The chronicles of cognition. Think-aloud and its potential in research on learning and instruction]
Status: revise and resubmit (major revision)

Autor*in	Position	wissenschaftliche Ideen (in %)	Datenerhebung und -aufbereitung (in %)	Datenanalyse und -interpretation (in %)	Manuskripterstellung und -überarbeitung (in %)
Sarah Bez	1	80	100	80	85
Fabian Burkart	2	0	0	10	5
Martin J. Tomasik	3	5	0	0	5
Samuel Merk	4	15	0	10	5

Artikel 3: Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers

Wurster, S., Bez, S., & Merk, S. (2023). Does learning how to use data mean being motivated to use it? Effects of a data use intervention on data literacy and motivational beliefs of pre-service teachers. *Learning and Instruction*, *88*, 101806. https://doi.org/10.1016/j.learninstruc.2023.101806
Status: Publiziert

Autor*in	Position	wissenschaftliche Ideen (in %)	Datenerhebung und -aufbereitung (in %)	Datenanalyse und -interpretation (in %)	Manuskripterstellung und -überarbeitung (in %)
Sebastian Wurster	1	50	90	50	50
Sarah Bez	2	10	0	10	30
Samuel Merk	3	40	10	40	20

Artikel 4: Data-based decision making in einer digitalen Welt: Data Literacy von Lehrpersonen als notwendige Voraussetzung

Bez, S., Tomasik, M. J., & Merk, S. (2023). Data-based decision making in einer digitalen Welt: Data Literacy von Lehrpersonen als notwendige Voraussetzung. In K. Scheiter & I. Gogolin (Hrsg.), *Bildung für eine digitale Zukunft* (S. 339–362). Springer VS. https://doi.org/10.1007/978-3-658-37895-0_14 [Edition ZfE]
Status: Publiziert

Autor*in	Position	wissenschaftliche Ideen (in %)	Datenerhebung und -aufbereitung (in %)	Datenanalyse und -interpretation (in %)	Manuskripterstellung und -überarbeitung (in %)
Sarah Bez	1	80	-	-	85
Martin J. Tomasik	2	10	-	-	5
Samuel Merk	3	10	-	-	10

Appendix

Appendix A: Graph types

Exemplary screenshots of graph types of the technology-based formative assessment system to date of the think-aloud-sessions.

Figure A.1

Graph 1: Learning progress on class or individual level

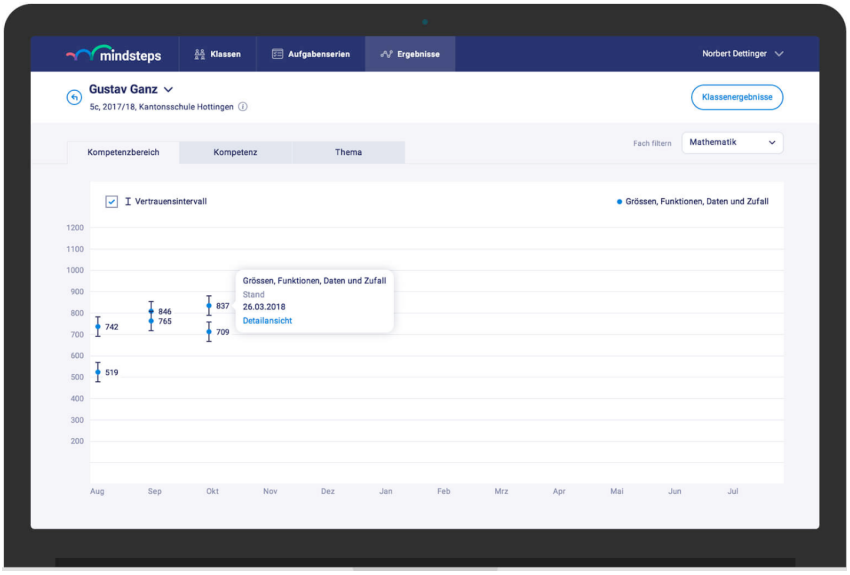


Figure A.2

Graph 2: class overview

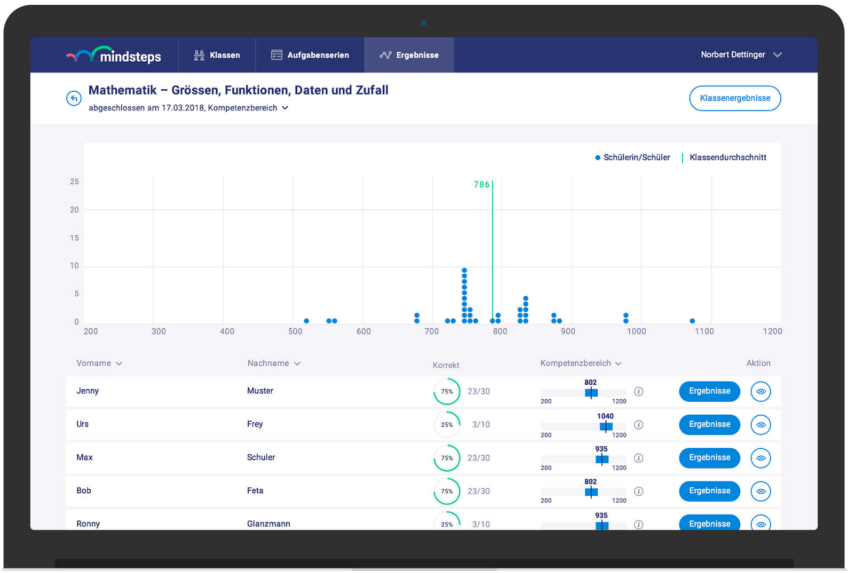


Figure A.3

Graph 3: tasks and task answers of individual students

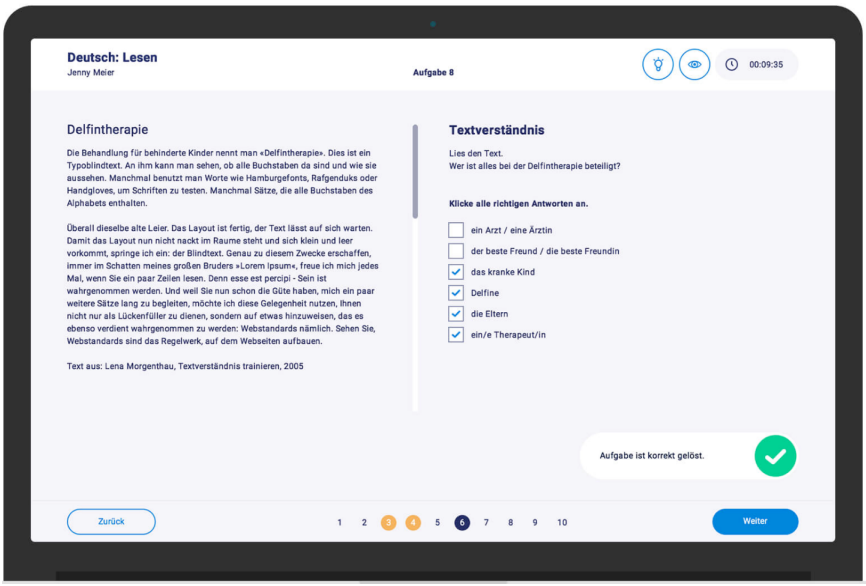


Figure A.4

Graph 4: competence levels

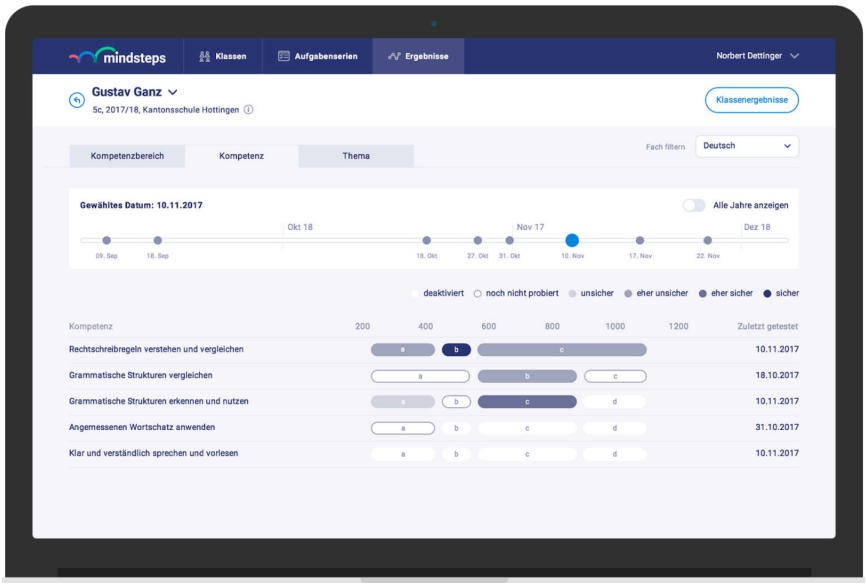
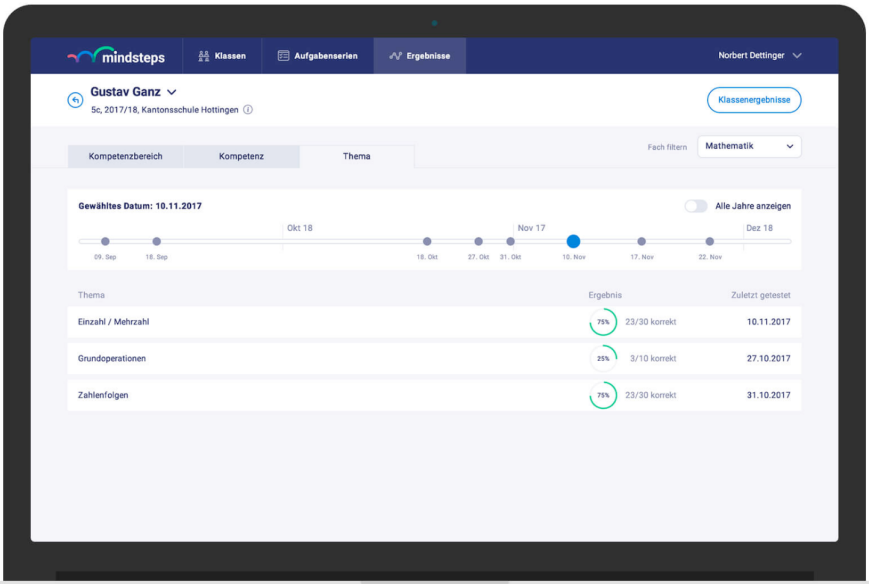


Figure A.5

Graph 5: overview of specific curriculum topics



Appendix B: Co-occurence of sensemaking steps and relative durations of addressing graph types

Figure B.1

Relative durations of co-occurrence of sensemaking steps

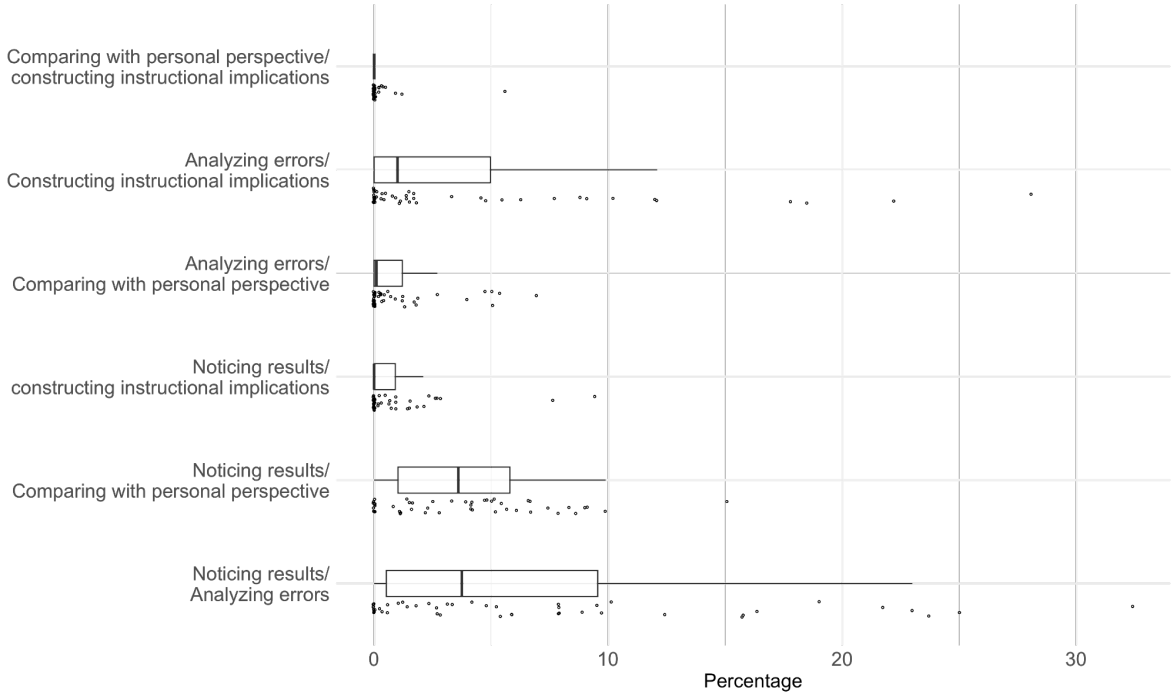
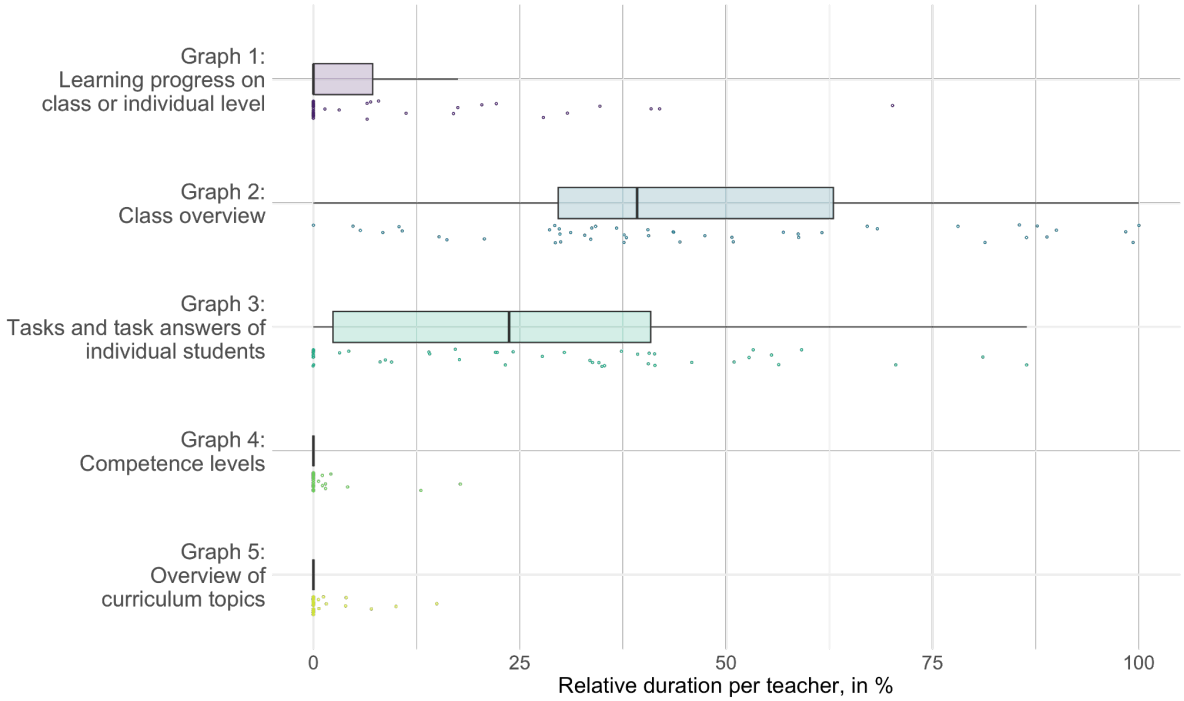


Figure B.2

Relative durations of addressing different graph types



Appendix Artikel 3

Table A1

Components of data literacy training (Filderman et al., 2021; Mandinach & Gummer, 2016) and corresponding intervention content

Components of data literacy training	Intervention content	Examples of learning activity
Identify problems and frame questions	N/A	N/A
Use data (location)	*Introduction in the data literacy module* *1) Data use in school-based instruction and data collection* - Introduction to the basic ideas of DBDM - Overview of different data sources - Basics of generating data and measuring variables - Scale levels - Levels of inference - Causal relationships - Validity, reliability and objectivity - Tips for applying data use in practice	- Learning video and presentation slides - Exercise about data collection - Quiz with feedback - PDF file with summary of important content and exercise task

(Continued on next page)

Table A1 (continued)

Components of data literacy training (Filderman et al., 2021; Mandinach & Gummer, 2016) and corresponding intervention content

Components of data literacy training	**Intervention content**	**Examples of learning activity**
Transform data into information (comprehension)	*2) Data reduction and data interpretation* Introduces the description of distributions of variables a) using quantitative characteristics, b) using qualitative terms and c) graphically - Characteristics of frequency distributions: central tendency, dispersion, modality, skewness - Graphical displays of frequency distributions: histograms, jitter plots, box plots, violin plots - Relating topics to data use processes	- Learning video and presentation slides - Interactively explore the ability of the different visualizations to depict central tendency, dispersion and modality skewness using two interactive shiny apps - Peer feedback for an exercise about data interpretation of the distribution of grades in different courses - Elaborations and exercises about descriptive distributions - Quiz with feedback

(continued on next page)

Table A1 (continued)

Components of data literacy training (Filderman et al., 2021; Mandinach & Gummer, 2016) and corresponding intervention content

Components of data literacy training	Intervention content	Examples of learning activity
Transform data into information (comprehension) (continued)	*3) Data transformation and data reduction* Introduction to techniques of - Data transformation (scaling, centring, standardization, percentiles, z-standardisation) - Data reduction (comparison of means and correlation of variables) - Judging mean differences using common language effect sizes - Relating topics to data use processes	- Learning video and presentation slides - Interactively explore z-values (guessing z-values), different meanings of z-values and percentiles and the comparison of means applying effect sizes and different visualizations using three interactive shiny apps - Worked out example and exercise using a case study about student feedback on instruction, including all steps of data use - Quiz with feedback
Transform information into a decision (Interpretation)	Includes 3) Data transformation and data reduction (see above)	
Evaluate outcomes	N/A	N/A

Table A2

Factor loadings (exploratory factor analysis) of the cluster robust three-factor model

	Factor one: Identify problems and frame questions & use data	**Factor two: Transform data into information**	**Factor three: Transform information into a decision and evaluate outcomes**
Item 1	**0.589**	0.049	0.195
Item 2	**0.285**	**0.412**	0.227
Item 3	**0.809**	0.205	-0.009
Item 4	**0.741**	-0.022	0.184
Item 5	-0.043	**0.881**	0.001
Item 6	-0.006	**0.482**	0.391
Item 7	-0.033	**0.684**	0.212
Item 8	0.016	**0.920**	-0.239
Item 9	0.060	**0.642**	0.126
Item 10	-0.169	0.043	**0.811**
Item 11	0.033	0.013	**0.780**
Item 12	0.016	-0.226	**1.003**
Item 13	-0.121	-0.012	**0.834**

Table A3

Reliability ω (McDonald, 1999) motivational beliefs about DBDM of the cluster robust three factor model (Study 1)

	Factor one: Identify problems and frame questions & use data	**Factor two: Transform data into information**	**Factor three: Transform information into a decision and evaluate outcomes**
Cost	.86	.90	.93
Enjoyment	.84	.92	.88
Self-efficacy	.86	.92	.89
Utility	.84	.88	.77
Attainment	.84	.87	.83

Table A4

Reliability ω (McDonald, 1999) motivational beliefs about DBDM scales (Study 2)

	Factor one: Identify problems and frame questions & use data	**Factor two: Transform data into information**	**Factor three: Transform information into a decision and evaluate outcomes**
Cost	.85	.91	.91
Enjoyment	.84	.90	.93
Self-efficacy	.89	.92	.92
Utility	.80	.88	.90
Attainment	.82	.85	.88

Table A5

Results of the multi-group trivariate latent change score models of motivational beliefs and the univariate latent change score model of the data literacy test.

Scale	Group	Standardized latent change scores			Posterior model probabilities of Bayesian hypothesis testing		
		α_1	α_2	α_3	H1	H2	H3
Data literacy test	TG	0.76			0.000	0.818	0.182
	CG	0.14					
Cost	TG	-0.03	-0.02	0	0.992	0.008	0.000
	CG	-0.02	0.11	0.13			
Enjoyment	TG	0.01	0.19	-0.02	0.930	0.063	0.007
	CG	-0.02	-0.23	-0.02			
Self-efficacy	TG	0.72	0.60	0.31	0.000	0.010	0.990
	CG	0.42	0.19	0.43			
Utility	TG	0.47	0.06	0.16	0.011	0.698	0.291
	CG	0.23	0.02	0.01			
Attainment	TG	0.31	0.11	0.04	0.595	0.365	0.040
	CG	0.13	-0.06	0.03			

Notes. TG = treatment group, CG = control group; α_1/ α_2/α_3 = latent difference scores factor one "Identify problems and frame questions & use data"/factor two "Transform data into information"/ factor three "Transform information into a decision & evaluate outcomes". PMPa = posterior model probabilities of three hypotheses: H1: CG α_1= α_2= α_3=0 and TG α_1= α_2= α_3=0; H2: CG α_1= α_2= α_3=0 and CG α_1 < TG α_1 & CG α_2 < TG α_2 & CG α_3 < TG α_3 ; H3: CG α_1 < TG α_1 & CG α_2 < TG α_2 & CG α_3 < TG α_3.