Bez, Sarah; Poindl, Simone; Bohl, Thorsten; Merk, Samuel

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

Zeitschrift für Pädagogik 67 (2021) 4, S. 551-572

D 7484

ZEITSCHRIFT FÜR PÄDAGOGIK

Heft 4 Juli/August 2021

■ *Thementeil*

Demokratieerziehung und die Herausforderungen des Liberalismus

■ *Allgemeiner Teil*

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

Eine Analyse über die Veränderung von Bildungsaspirationen von SchülerInnen nach dem Übergang in die Sekundarstufe

Kooperation zwischen Verstehen und Nichtverstehen – Systemtheoretische Modellierungen zur inklusionsbezogenen Kooperation von Lehrkräften unterschiedlicher Lehrämter im Schulunterricht

BELTZ JUVENTA

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 67 – Heft 4
Juli/August 2021

Inhaltsverzeichnis

Thementeil: Demokratieerziehung und die Herausforderungen des Liberalismus

Allgemeiner Teil

Besprechungen

Dokumentation

Table of Contents

Topic: Democratic Education and the Challenges to Liberalism

Articles

Allgemeiner Teil

Sarah Bez/Simone Poindl/Thorsten Bohl/Samuel Merk

Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert?

Ergebnisse zweier Think-Aloud-Studien

Zusammenfassung: 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.

Schlagworte: Vergleichsarbeiten, Rezeption, Datenkompetenz, Think-Aloud-Methode, datenbasierte Unterrichtsentwicklung

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, 2006, 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, Moosbrugger & Zuber, 2016). Übersichtsarbeiten, die die bisherige Forschung seit der Einführung von Vergleichsarbeiten zu deren Rezeption und Nutzung durch Lehrpersonen zusammen-

fassen, 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, Metz, Bohl, Kleinknecht & Schymala, 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, Richter & Lenski, 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 sind und inwiefern sich Zusammenhänge zur allgemeinen Datenkompetenz zeigen.

2. Theoretischer Hintergrund und Forschungsstand

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 knowledge, and an understanding of how children learn. (2016, S. 2)

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, Curcio & Bright, 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, Espin, Chung & Saab, 2017; van den Hurk, Houtveen & van de Grift, 2016; Zeuch, Förster & Souvignier, 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).

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, Hosenfeld & Koch, 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, Poindl, Wurster & Bohl, 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.

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, Maritzen, Eckert & Thiel, 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.

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, Wayman, Deno, McMaster & Rooij, 2017; Leighton, 2017; Padilla & Leighton, 2017; van Someren, Barnard & Sandberg, 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?

3. Methode[1]

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

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

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).

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 Abb. 1 und 2). Anschließend bearbeiteten die Studierenden einen Datenkompetenztest (siehe Abschnitt 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 3.4).

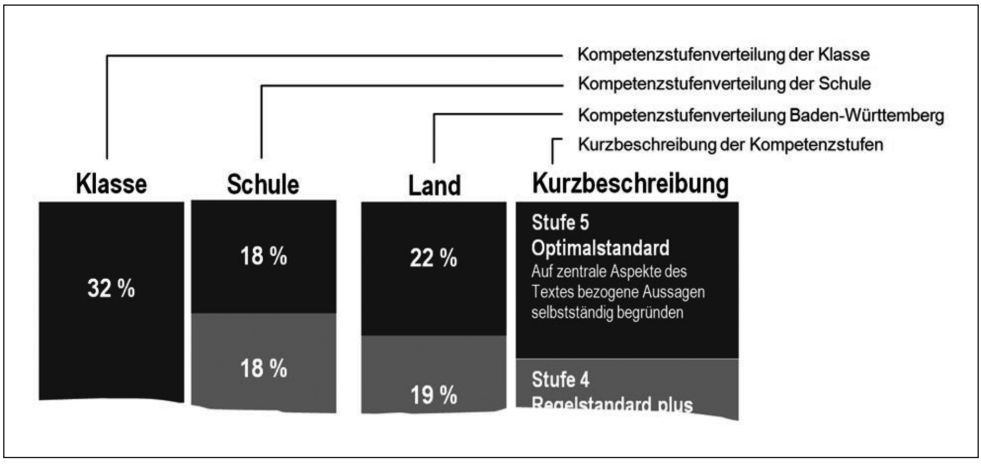


Abb. 1: Kompetenzstufengrafik (Ausschnitt), © Institut für Bildungsanalysen Baden-Württemberg (IBBW), 2019, S. 7, Abdruck mit freundlicher Genehmigung

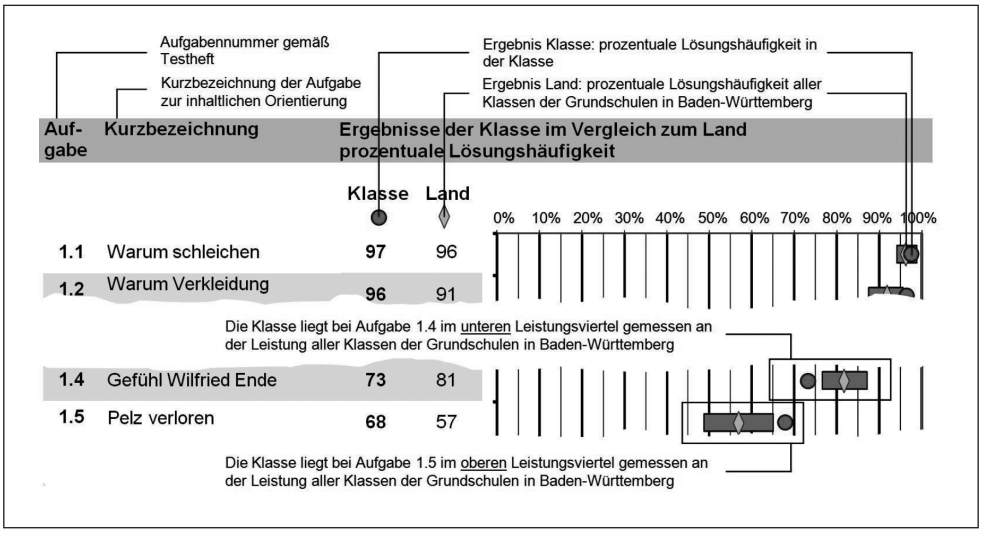


Abb. 2: Lösungshäufigkeiten der einzelnen Aufgaben (Ausschnitt), © Institut für Bildungsanalysen Baden-Württemberg (IBBW), 2019, S. 8, Abdruck mit freundlicher Genehmigung

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 3.4).

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, Guhn & Zumbo, 2012) und zeigte gute Ergebnisse (Studie 1: $\alpha = .767$; Studie 2: $\alpha = .737$).

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 Rück-

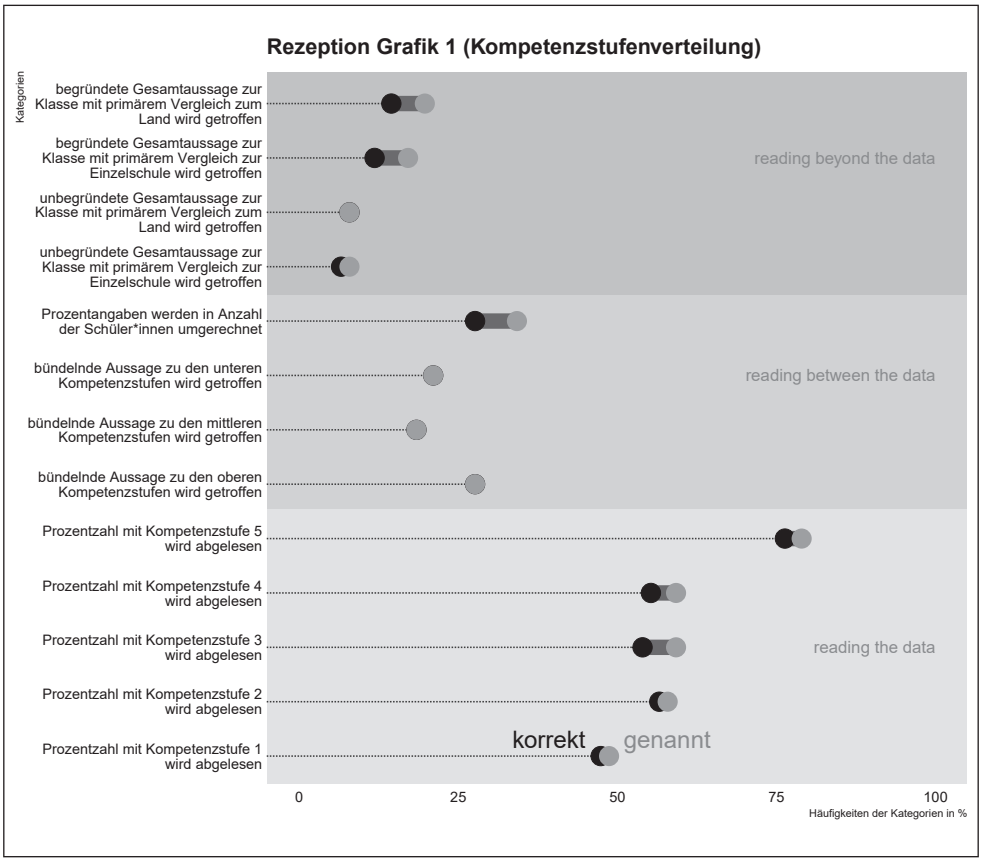
meldungen rezipierten, wurden nur diese weiter ausgewertet. Es handelte sich dabei (wie in Studie 1) um die Darstellungen der Kompetenzstufenverteilungen (Grafik 1) und der Lösungshäufigkeiten zu einzelnen Aufgaben (Grafik 2) zu den jeweils geprüften Kompetenzbereichen. Im Zuge einer zweiten induktiven 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 Interraterreliabiliät ergab befriedigende bis sehr gute Werte ($.58 \leq \alpha \leq .90$).

4. Ergebnisse

4.1 Studie 1: Lehramtsstudierende

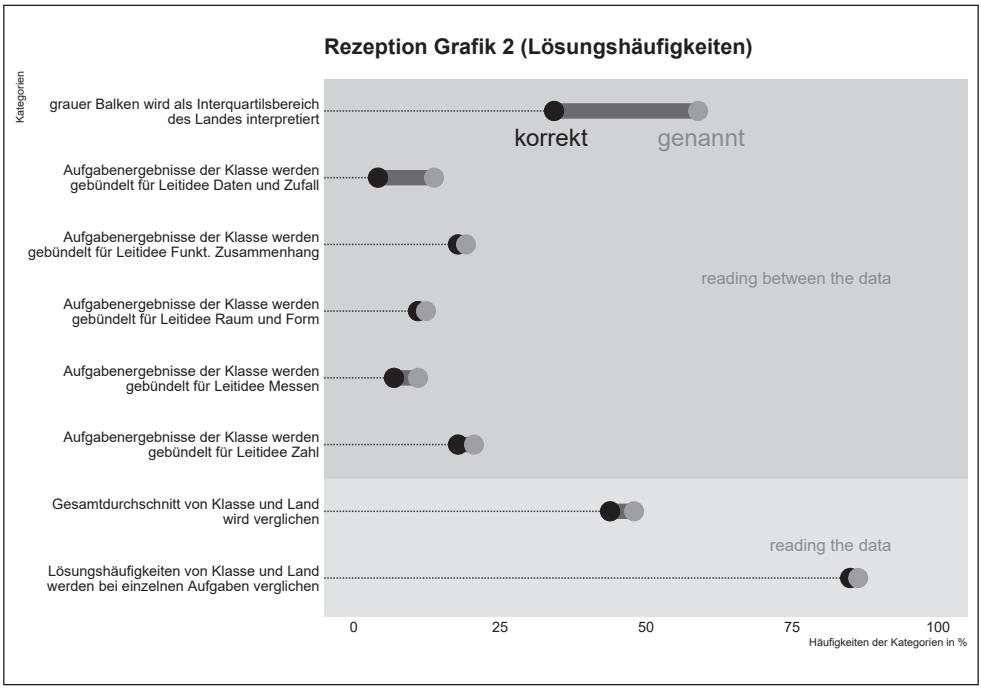
Forschungsfrage 1
Auf Grundlage der Think-Aloud-Kodierungen in Studie 1 wurde bezüglich der ersten Forschungsfrage untersucht, welche Informationen Lehramtsstudierende den Grafiken entnehmen und welche komplexeren Elaborationen (wie etwa die Gruppierung von Daten) vorgenommen werden. Dabei wurde auch 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 (Abb. 3: Rezeption Grafik 1 [Kompetenzstufenverteilung]; Abb. 4: Rezeption Grafik 2 [Lösungshäufigkeiten]).

Abbildungen 3 und 4 zeigen, 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 % und mehr 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.



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/.

Abb. 3: Rezeption Grafik 1 (siehe Abb. 1)



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/.

Abb. 4: Rezeption Grafik 2 (siehe Abb. 2)

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 (β_1 = .586*, p = .043) sowie für die korrekte Interpretation des Interquartilsbereich des Landes (β_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, Mulder & Hoijtink, 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 sind als unter der Annahme des Gegenteils ($\beta_1 \leq 0$), was meist als substantielle Evidenz interpretiert wird. Bei *reading beyond the data* bei Grafik 1 zeigten sich inkonklusive Ergebnisse ($BF_{1c} = .615$).

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 1 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.

	grafikübergreifend	**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]

Anmerkung. Angegeben sind jeweils Median sowie 1. und 3. Quartil der prozentualen Anteile bezogen auf die Gesamtdauer des Think-Aloud-Protokolls.
Bsp. *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 %.

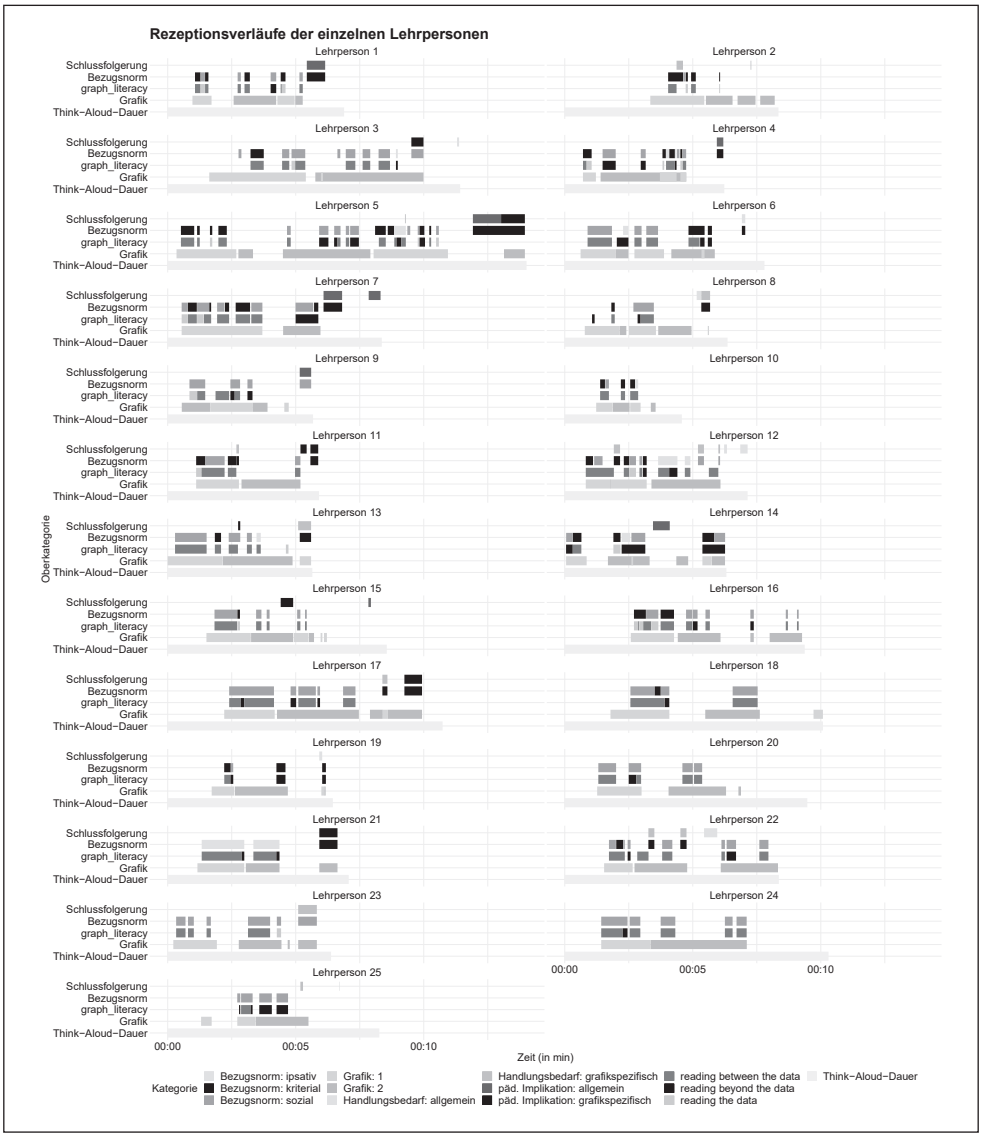
Tab. 1: Anteile der graph literacy-Stufen und der Bezugsnormen bei der Rezeption

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 (Abb. 5).

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 (Abb. 5) 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, Ly, Marsman & Wagenmakers, 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 ($BF_{10} = 8.596$).



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/.

Abb. 5: Rezeptionsverläufe der einzelnen Lehrpersonen

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:* $BF_{10} = .499$).

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. Abb. 3 und 4). 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. Abb. 5). 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. 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 in den Think-Aloud-Protokollen 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, Poortman, Schildkamp & Pieters, 2017; Koch, 2013; van Geel, Keuning, Visscher & Fox, 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 kri-

terialen (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.

Literatur

Altrichter, H., & Maag Merki, K. (Hrsg.) (2016). *Handbuch Neue Steuerung im Schulsystem* (2. Aufl.). Wiesbaden: Springer.

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). Wiesbaden: Springer.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences.* New York: Cambridge University Press.

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.

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

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

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.

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.

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.

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.

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.

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.

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). Bern: hep.

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 [26. 03. 2021].

KMK (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* Köln: Wolters Kluwer.

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 [26. 03. 2021].

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

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

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). Wiesbaden: Springer VS.

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.

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

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.

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). Wiesbaden: Springer.

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). Münster: Waxmann.

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

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

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

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.

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

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). Wiesbaden: Springer Fachmedien.

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 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 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. 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 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.

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.

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). Münster/New York: Waxmann.

Abstract: Data analysis is considered a decisive step in the entire cyclic process of data-based decision-making; however, it is poorly understood. By conducting think-aloud interviews, this study examines how pre-service and in-service teachers analyze the results of statewide assessments. Results reveal that both pre-service and in-service teachers mainly noticed and compared the information that was directly displayed in graphs but rarely showed elaborations with higher complexity. In addition, social comparisons were the dominating reference norm. An association between generic data literacy and the complexity of elaborations did not consistently emerge. Furthermore, Bayes factors provide some evidence against the consistency between data analysis and pedagogical implications with regard to different graph types and reference norms.

Keywords: Statewide Assessments, Analyzing, Data Literacy, Think-aloud Method, Data-based Decision-making

Anschrift der Autor_innen

Sarah Bez, Eberhard Karls Universität Tübingen,
Institut für Erziehungswissenschaft, Abt. Schulpädagogik,
Münzgasse 22–30, 72070 Tübingen, Deutschland
E-Mail: sarah.bez@uni-tuebingen.de

Simone Poindl, Eberhard Karls Universität Tübingen,
Institut für Erziehungswissenschaft, Abt. Schulpädagogik,
Münzgasse 22–30, 72070 Tübingen, Deutschland
E-Mail: simone.poindl@uni-tuebingen.de

Prof. Dr. Thorsten Bohl, Eberhard Karls Universität Tübingen,
Institut für Erziehungswissenschaft, Abt. Schulpädagogik,
Münzgasse 22–30, 72070 Tübingen, Deutschland
E-Mail: thorsten.bohl@uni-tuebingen.de

Jun.-Prof. Dr. Samuel Merk, Pädagogische Hochschule Karlsruhe,
Institut für Schul- und Unterrichtsentwicklung in der Primar- und Sekundarstufe,
Bismarckstraße 10, 76133 Karlsruhe, Deutschland
E-Mail: samuel.merk@ph-karlsruhe.de