Der Nutzen von Kompetenzstufenmodellen im Rahmen datengestützter Unterrichtsentwicklung

Dissertation

zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)
im Fach Erziehungswissenschaften

eingereicht am: 17.11.2021
verteidigt am: 15.06.2022

an der Kultur-, Sozial- und Bildungswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von Peter Harych

Prof. Dr. Peter Frensch
(komm.) Präsident der
Humboldt-Universität zu Berlin

Prof. Dr. Christian Kassung
Dekan der Kultur-, Sozial- und
Bildungswissenschaftlichen Fakultät

Gutachter/in:
1. Prof. Dr. Anand Pant
2. Prof. Dr. Bettina Hannover

Danksagung

Diese Arbeit ist auch Fazit der zurückliegenden 15 Jahre Arbeit im ISQ und Dank gebührt deshalb einigen, deren Unterstützung ich dabei erfahren durfte. Ich danke Anand Pant für die Motivation und dass er mir immer das Gefühl gab, dass ich es schaffen kann. Für unterstützende Worte und die bereitwillige Übernahme des Gutachtens danke ich herzlichst Bettina Hannover. Zudem bin ich froh, dass mich mein Freund Rico Emmrich bei allen cleveren und besonders den weniger cleveren Gedanken kritisch begleitet hat und wünsche mir, dass er dies weiterhin tun möge. Sehr Danken möchte ich auch allen Menschen aus der Schulpraxis, die mit mir bereitwillig ihre Probleme geteilt und mir von ihrer aufreibenden aber oft auch gelingenden Arbeit erzählt haben, allen voran natürlich Gundula Meiering.

Einen späten Dank möchte ich an zwei besondere Menschen senden, die mir ein Vorbild sind und die mich mit der ihnen je eigenen Großzügigkeit in die (Bildungs-)Forschung hineingezogen haben, wie sonst niemand. Danke Anna Maciel und Rainer Peek.

Nicht zuletzt brauchte und brauche ich meine Familie. Claudia, Lotte, Jakob und Leonie - einen grandiosen Dank für Eure Unterstützung! Diese Arbeit ist für Dich Inge.

Zusammenfassung

Als Konsequenz der besonderen Aufmerksamkeit nach der Veröffentlichung der Ergebnisse des ersten Programme for International Student Assessment (PISA) (Baumert et al., 2001), wurden auf unterschiedlichen Ebenen des deutschen Schulsystems evaluatorische Prozesse implementiert. Zurückgegriffen wurde dazu auf schon andernorts bewährte Methoden, wie der empirischen Messung von Erträgen des Bildungssystems. In Deutschland fußen solche Messungen auf von der Konferenz aller Kultusminister (KMK) beschlossenen Bildungsstandards, die als Kompetenzstrukturmodelle entwickelt und als Kompetenzstufenmodelle für die Messung operationalisiert wurden. Mit den Vergleichsarbeiten etablierte sich in Deutschland ein Verfahren, das an diese Messung von Kompetenzerträgen anschließt, die Ergebnisse aber primär schulischen Rezipienten zur Verfügung stellt. Damit sich der intendierte Nutzen von datengestützter Unterrichts- und Schulentwicklung realisiert, muss Validität im Sinne Kanes (Kane, 2013) vorausgesetzt werden. Die vorliegende Arbeit untersucht dazu drei Aspekte des Validitätsarguments für Vergleichsarbeiten.

Dies sind zum Ersten mit der Nutzungsperspektive eng verbundene psychometrische Anforderungen an das Instrument. Das vierte Kapitel untersucht dazu methodische Aspekte, welche als grundlegend für die Validität angesehen werden, wobei Validität als Qualitätskriterium hinsichtlich der Zuverlässigkeit der Testwertinterpretationen (Hartig et al., 2020, S. 530) verstanden wird. Es werden Gewissheiten formuliert und in Frage gestellt, die für die Testkonstruktion, aber insbesondere für die Weiterarbeit mit den Testergebnissen in der Schule essentiell sind. Der Validität von Testwertinterpretationen über mehr als einen Messzeitpunkt widmet sich das fünfte Kapitel. Wegen der Art und Weise der Verknüpfung der Messskalen haben die Ergebnisse letztlich aber Relevanz für die Interpretation jeder Einzelmessung. Der dritte Aspekt von Nutzungsvoraussetzung wird im wohl meistzitierten Nutzungsmodell für die Vergleichsarbeiten von Helmke (2004) als *technische Übermittlung* der Rezeption untergeordnet und ist ein Element konsequenzieller Validität. Der Abruf von Rückmeldungen ist so trivial wie grundlegend und doch findet sich dazu (zumindest im deutschsprachigen Raum) nur eine einzige Veröffentlichung mit relevanten Daten. Die vorliegende Untersuchung ermöglicht erstmals einen dezidierten Einblick in diesen Aspekt der Nutzung von Vergleichsarbeiten.

Abstract

As a consequence of the special attention following the release of the results of the first Programme for International Student Assessment (PISA) (Baumert et al., 2001), evaluative processes were implemented at multiple different levels of the German educational system. There, previously proven methods, such as the empirical measurement of results of the educational system, were used. In Germany, these measurements are based on educational standards determined by the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK), which were developed as competence structure models and are being operationalized for the measurements as competence level models. With the Vergleichsarbeiten, a process was established in Germany that continues these measurements of competency outcomes, but whose results are primarily made available to recipients in schools. To realize the intended use of data-based lesson as well as school development, validity in Kane's sense (Kane, 2013) is required. For that reason, this paper examines three aspects of the validity argument for the Vergleichsarbeiten.

On one hand, these are psychometric requirements for the instrument, which are highly dependend on the intended use. In this regard, the fourth chapter examines methodical aspects seen as fundamental for validity, where validity is understood as a quality criterion in regards to the reliability of interpretations of test results (American Educational Research Association (AERA) et al., 2014, p. 11). Certainties are defined and questioned, that are essential for test construction, but also especially for working with the test results in lesson development. The fifth chapter is concerned with the validity of test result interpretations spanning multiple measurement times, but because of the way the scales of measurement are connected, the results end up being relevant to the interpretation of singular measurements as well. The third aspect is an element of consequentional validity which, as *technical transmission*, is subordinated to reception, in the most cited usage model for the Vergleichsarbeiten by Helmke (2004). The download of Feedback is as trivial as it is fundamental, yet (at least in the German-speaking world) only a single publication with relevant data can be found. This examination allows, for the first time, a dedicated look into this aspect of the Vergleicharbeiten's usage.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

2/3-PL	2/3-Parameter Logistic (Modell)
AIC	Akaike Information Criterion (Informationsmaß)
BiSta	Bildungsstandards
CI	Confidence Intervall (Konfidenz- oder Vertrauensintervall)
CPU	Central Processor Unit (zentraler Prozessor in einem Computer)
DeSeCo	Definition and Selection of Competencies
GemS	Gemeinschaftsschule
GER	Gemeinsamer Europäischer Referenzrahmen für Sprachen
GEW	Gewerkschaft Erziehung und Wissenschaft
GYM	Gymnasium
HSA	Hauptschulabschluss
ICC	Item Characteristic Curve
ICD-10	International Statistical Classification of Diseases and Related Health Problems (Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme) in der aktuell gültigen Version 10
ID	Identification (Synonym für einen eine Entität eindeutig identifizierende Variable)
IQB	Institut zur Qualitätsentwicklung im Bildungswesen
IRT	Item Respone Theory
ISQ	Institut für Schulqualität der Länder Berlin und Brandenburg e.V.
ISS	Integrierte Sekundarschule
JbL	Jahrgangsbezogene Lerngruppe
JüL	Jahrgangsübergreifende Lerngruppe
KERMIT	Hamburger Tests zur Lernstandsfeststellung und -entwicklung „Kompetenzen ermitteln“ (in den Klassen 2, 3, 5, 7, 8 und 9)
KTT	Klassische Testtheorie
LSA	Large-Scale-Assessment

KMK	Kultusministerkonferenz
MBJS	Ministerium für Bildung, Jugend und Sport (des Landes Brandenburg)
MSA	Mittlerer Schulabschluss
NEPS	Nationales Bildungspanel (National Educational Panel Study)
NKLM	Nationaler Kompetenzbasierter Lernzielkatalog Medizin
OECD	Organisation for Economic Cooperation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
OIB	ordered item booklet (Handbuch mit nach Schwierigkeit geordneten Items)
OR	Odd Ratio (Chancenverhältnis)
PC	Personal Computer
PDF	Portable Document Format (transportables Dokumentformat)
PISA	Programme for International Student Assessment
PISA-E	PISA-Erweiterung (nationale Ergänzung der internationalen PISA-Studien)
PV	plausible values
RAM	Random-Access Memory (Speicher mit direktem Zugriff, i.A. Arbeitsspeicher)
SenBJF	Senatsverwaltung für Bildung, Jugend und Familie (des Landes Berlin)
TIMSS	Trends in International Mathematics and Science Study
ÜGK	Überprüfung der Grundkompetenzen (Die Überprüfung der nationalen Bildungsziele für einige Bereiche der obligatorischen Schule in der Schweiz wird als Erhebung zur ÜGK bezeichnet und ist damit das Äquivalent zum IQB-Bildungstrend in Deutschland.)
VERA	Vergleichsarbeiten
VERAMSA	Berliner Studie zur Untersuchung der Leistungsentwicklung von VERA 8 (Klasse 8) bis zum Mittleren Schulabschluss (Klasse 10)
WLE	Weighted-Likelihood-Estimates
ZENSOS	Zentrales System zur Online-Verwaltung von Schulinformationen

Einleitung

Etwa Mitte der 1990er Jahre wurde Bildungsforschung und Qualitätssicherung im Bildungswesen auch im deutschsprachigen Raum vermehrt mit quantitativen empirischen Methoden betrieben (zum Beispiel Lehmann et al., 1996; Baumert & Lehmann, 1997). Die Auswertung der Ergebnisse des *Programme for International Student Assessment* (PISA) des Jahres 2000 (Baumert et al., 2001) verschafften diesen Aktivitäten wohl insbesondere wegen des unerwartet schlechten Abschneidens Deutschlands eine herausgehobene Aufmerksamkeit. Maritzen (2014) sieht diese besondere Aufmerksamkeit der Öffentlichkeit schon 1996 nach den TIMSS-Ergebnissen (Baumert & Lehmann, 1997) und beschreibt in seinem Aufsatz, dass dieser Aufbruch zumindest von politischer Seite aus mehr einer „spezifisch historischen Konstellation“ zuzuschreiben ist, denn einer gezielten Entscheidung für evidenzbasierte Qualitätsentwicklung. Als Konsequenz wurden dann aber tatsächlich auf unterschiedlichen Ebenen des Schulsystems evaluatorische Prozesse implementiert und dazu zentral auf die empirische Messungen von Erträgen des Bildungssystems und deren Darstellung in Form von Kompetenzen zurückgegriffen.

Die Messung fachlicher Kompetenzen erfolgt verbunden mit konkreten Zielstellungen auf unterschiedlichen Ebenen und mit den zentralen Vergleichsarbeiten letztlich auch im Rahmen schulischer und unterrichtlicher Qualitätsentwicklung. Weinert (2001a, S.30) formuliert als Anforderungen an schulische Leistungsmessungen, dass diese methodisch zuverlässig, sensibel für variable Kontextbedingungen sein und sich auf die eigentliche pädagogische Fragestellung beziehen müssen. Vergleichsarbeiten sind dabei vielfältige Zielstellungen zugeordnet worden, wobei aber Unterrichtsentwicklung bis heute in jedem Fall als prioritär benannt wird (KMK, 2010, S.17; KMK, 2012a; KMK, 2018b). Die vorliegenden Untersuchungen widmen sich speziell der Nutzung der Ergebnisse von Vergleichsarbeiten für die Unterrichtsentwicklung. In Frage steht hierbei, inwieweit sich die Vergleichsarbeiten mit ihren Kompetenzstufenmodellen als zentral entwickeltes und in einer top-down-Strategie in allen Ländern implementiertes Instrument, für die Lösung lokaler Problemstellungen als geeignet erweisen und so genutzt werden.

Vorgängige Untersuchungen zur Nutzung der Vergleichsarbeiten wie beispielsweise von Koch (2011), Groß Ophoff (2013) oder Schliesing (2017) untersuchen konkrete Prozesse der Auseinandersetzung von Rezipienten mit den Rückmeldungen von Vergleichsarbeiten, verknüpften dazu qualitative und quantitative, teilweise quasi-experimentelle Methoden und untersuchen den Einfluss von Merkmalen wie Verständlichkeit, Nützlichkeit und Akzeptanz auf die tatsächlich stattfindende Nutzung. Sie rekurrieren dabei auf verschiedene Modelle der Ergebnisnutzung. Eine umfassende Übersicht dieser liefert Pukrop (2019). Die vorliegende Arbeit fokussiert ergänzend dazu drei für die Interpretation und Nutzung kompetenz(stufen)bezogener Rückmeldungen von Vergleichsarbeiten grundlegende Aspekte, die hier als methodisch und praktisch vorgelagert interpretiert werden. Diese lassen sich grob drei Kernfragen zuordnen.

1. Inwiefern unterstützen Schlussfolgerungen aus Modellannahmen der Rasch-Skalierung, die hier Gewissheiten genannt werden, die Interpretation von Fähigkeitsparametern?

Im Kapitel *Überprüfung von Gewissheiten beim Einsatz der Rasch-Skalierung* wird mit Bezug auf Hartig et al. (2020) die Übertragung des sich in Large-Scale-Assessments bewährten Messkonzepts auf die Vergleichsarbeiten verbunden mit der Fragestellung untersucht, inwieweit dem potentiell möglichen Nutzen für Unterrichtsentwicklung Validität bescheinigt werden kann. Nach einem kurzen Problemaufriss werden vier Hypothesen formuliert, die insbesondere für die Weiterarbeit mit den Testergebnissen im Unterricht essentiell sind. Hinterfragt wird im Einzelnen (1) die Irrelevanz der Itemauswahl, (2) die erwartungstreue Schätzung von Personenparametern, (3) die Bedeutung von Guttman-Pattern sowie (4) der Zusammenhang von Itemschwierigkeiten und Personenfähigkeiten. An die Darstellung der Methoden und Daten schließt sich die Präsentation der Ergebnisse der Hypothesenprüfung und eine Zusammenfassung in tradierter Weise an. Ein kurzes Resümee zur technischen Realisierung der Simulationen schließt dieses Kapitel ab.

2. Erweisen sich Interpretationen der Ergebnisse von Vergleichsarbeiten aus unterschiedlichen Jahren auf der gemeinsamen Skala der Bildungsstandards als valide?

Der Validität von Testwertinterpretationen über mehr als einen Messzeitpunkt widmet sich das anschließende Kapitel *Stabilität der Ergebnisse von Vergleichsarbeiten.* Auch wenn die Standardrückmeldungen für Rezipienten in Schulen solche Interpretationen in der Regel nicht vorsehen, so erweisen sich Trendaussagen im Rahmen von Rechenschaftslegung auf der Ebene einzelner Schulen oder auch auf höheren Aggregationsebenen als wirkmächtig und der

Nachweis von Validität damit als immanent. Insbesondere deshalb, weil diese Interpretation mindestens bei der Implementation der Vergleichsarbeiten intendiert war. Wie gezeigt wird, haben diese Ergebnisse wegen der zentralen Bedeutung der Kompetenzstufenmodelle letztlich aber Relevanz für die Interpretation auch jeder Einzelmessung. Dieses Kapitel gliedert sich in zwei Teile, die in unterschiedlicher Weise eine Untersuchung der Messstabilität vornehmen. Nachdem der Begriff Stabilität definiert worden ist, werden im ersten Teil drei artefakte Ergebniskonstellationen aus verschiedenen VERA-Kohorten präsentiert und deren Plausibilität diskutiert. Der zweite Teil enthält hingegen zwei Analysen, bei denen jeweils VERA-Daten zu einem Längsschnitt ergänzt und ausgewertet werden. Im ersten Fall werden Daten einer 2011 gestarteten Panelstudie unter dem hier zentralen Gesichtspunkt der Stabilität bewertet, wobei bisher ungenutzte Daten aus zwei Messzeitpunkten einbezogen werden. Die Panelstudie nutzt dabei VERA-Instrumente aus unterschiedlichen Jahren in verschiedenen Konstellationen. Die zweite Studie untersucht die Wiederholung einer Testdomäne nach einem Jahr mit einem identischen VERA-Instrument zur Feststellung des Zugewinns. Diese zwei Untersuchungen sind als Studien mit jeweils eigenem Methoden- und Ergebnisteil in das Kapitel eingebettet.

3. Können die Abrufe der Rückmeldungen die Annahme gelingender Unterrichts- und Schulentwicklung als Konsequenz der Vergleichsarbeiten bestätigen?

Für die Ergebnisnutzung von ebenso grundlegender Bedeutung ist der im Kapitel *Vor der Rezeption* untersuchte reale Zugriff auf die zur Verfügung gestellten Rückmeldungen. Viele Autoren (Groß Ophoff, 2013; Stamm, 2003; Hosenfeld, 2010; Hartung-Beck, 2009) finden bei Ihren Untersuchungen der Nutzung von leistungsbezogenen Rückmeldungen verschiedene Reflexionstypen. Gemein ist allen Modellen, dass sich Rezipienten zu einer Rückmeldung positionieren. In keinem Modell findet sich eine Gruppe von nicht-Rezipienten und somit keine Abschätzung zu ihrem Anteil. Die vorliegende Untersuchung operationalisiert mit dem Zugriff auf Rückmeldungen eine Grundlage (fast) jeder Ergebnisnutzung. Ein erfolgter Abruf gibt aber keine Auskunft über Zustimmung oder Ablehnung von Vergleichsarbeiten oder gar über deren Gründe. Das Rückmeldeabrufe als Grundbedingung für die Rezeption bisher kaum eine Rolle spielten kann daran liegen, dass die überwiegende Zahl der Rückmeldungen auch abgerufen wird. Diese Untersuchung ist ein Beitrag dazu, diese Wissenslücke zu schließen. Validitätstheoretisch ist das Merkmal des Rückmeldeabrufs ein Qualifikator für Schlussfolgerungen der Interpretation und Nutzung (siehe Abschnitt 2.3), für bestimmte Zielstellungen der Vergleichsarbeiten aber auch ein die konsequentielle Validität stützendes Argument. Die

Präsentation des Forschungsstandes zeigt, dass mit dieser Untersuchung erstmals ein Einblick in solche Prozesse offeriert wird. Für bisherige Arbeiten lassen sich bis auf eine einzige Ausnahme nur implizite Bezüge herstellen. Auf Grund der damit fehlenden theoretischen Grundlagen werden Forschungsfragen statt Hypothesen formuliert, dann aber klassisch bearbeitet.

Im ersten Kapitel *Die Vergleichsarbeiten als Instrument der Schul- und Unterrichtsentwicklung* werden die Vergleichsarbeiten als zentraler Gegenstand dieser Arbeit eingeführt. Das dieser Arbeit zugrundeliegende Konzept von Validität wird im zweiten Kapitel *Validität als integriertes, bewertendes Urteil* erläutert. Anschließend werden die drei Untersuchungen als Schlussfolgerungen eines Validitätsarguments nach Kane (2013) in einen gemeinsamen Rahmen eingebettet. Im Kapitel 3, *Raschskalierung, Standardsetting und Linking* werden die für das Verständnis der Untersuchungen notwendigen, vor allem psychometrischen Grundlagen zusammengefasst. Die drei zentralen Kapitel *Überprüfung von Gewissheiten beim Einsatz der Rasch-Skalierung* (4), *Stabilität der Ergebnisse von Vergleichsarbeiten* (5) sowie *Vor der Rezeption* (6) beginnen jeweils mit einer Problembeschreibung und fassen die Ergebnisse in einer Übersicht zusammen. Die Strukturen dieser drei Kapitel sind indes Problem-adäquat unterschiedlich. In der Gesamtdiskussion (Kapitel 7) werden die Ergebnisse der drei untersuchten Validitätsfacetten zusammenfassend bewertet, einige Forschungsdesiderata abgeleitet und Limitationen ergänzt.

1. Die Vergleichsarbeiten als Instrument der Schul- und Unterrichtsentwicklung

Die Vergleichsarbeiten (VERA) sind ein für Deutschland zentral entwickeltes Testinstrument, zur jährlichen Überprüfung von Kompetenzständen bei Schülerinnen und Schüler der dritten und achten Jahrgangsstufe. Die Testungen finden für alle Schüler*innen des achten Jahrgangs im Allgemeinen nach den Winterferien Ende Februar, Anfang März statt und im dritten Jahrgang zwischen Ende April und Anfang Mai. Für die von den einzelnen Ländern eigenständig administrierte Durchführung steht heute ein Testkorridor von jeweils mehreren Wochen zur Verfügung. In den ersten VERA-Jahren wurden bundesweit zentrale Testtermine festgelegt. Zur Durchführung haben sich die Kultusminister*innen aller Länder in einer Gesamtstrategie verpflichtet (zuerst KMK, 2006b; später KMK, 2016a).

In den folgenden Abschnitten werden zuerst die Vergleichsarbeiten in die Gesamtstrategie der Kultusministerkonferenz (KMK) zum Bildungsmonitoring eingeordnet. Im Weiteren wird die Genese der deutlich vor der Gesamtstrategie beginnenden Vergleichsarbeiten skizziert, die Entwicklung nachgezeichnet und ein Ausblick über die bevorstehenden Veränderungen gegeben. Diese Einordnungen sind relevant, weil sie die teilweise unscharfe Festlegung bzw. Interpretation der Zielstellung von VERA unter verschiedenen Perspektiven thematisieren. Die Nutzung der Ergebnissen, die in dieser Arbeit beleuchtet wird, ist eng mit dieser Zielstellung verbunden. Die Zielstellung für ein Instrument wie die Vergleichsarbeiten ergibt sich dabei nicht allein aus einer offiziellen Erläuterung, sondern ist eng verbunden mit Entscheidungen zu a) der Konstruktion des Tests und den sich damit ergebenden Möglichkeiten, b) der Administration der Testdurchführung sowie c) der Gestaltung der Rückmeldungen und d) der (zumeist implizit formulierten) Hoheit über die Daten und der damit verbundenen Entscheidung, welche Adressaten welche Informationen erhalten. Je mehr die Schulleistungsmessung im Weinert'schen Sinne (2001a) sensibel für variable Kontextbedingungen ist, desto mehr versperrt sie sich der Nutzung eines kontrollierendem, vergleichenden Monitorings. Auch deshalb werden solche Aspekte in den folgenden Ausführungen in Augenschein genommen.

Über eine Quellenanalyse im Abschnitt 1.3 wird die offizielle Zielbestimmung der Vergleichsarbeiten über den bisherigen Entwicklungszeitraum nachgezeichnet. Unstrittig und einheitlich in den Dokumenten der KMK wie den ländereigenen Strategien bzw. Verlautbarungen ist Unterrichts- und Schulentwicklung[1] die Kernfunktion der Vergleichsarbeiten. Die für die Länder darüber hinaus spezifische Ergebnisnutzung bezieht Elemente des Monitorings mehr oder weniger ein und kommuniziert diese Tatsache mehr oder weniger deutlich.

1.1. Kompetenzbegriff

Sicher muss auch der Klieme-Expertise (Klieme et al., 2003) zugeschrieben werden, dass nach PISA 2000 in Deutschland und im deutschen Sprachraum insgesamt die Entwicklung von Kompetenzen fokussiert wurde. Klieme et al. zitieren in ihrer Expertise die Definition des Kompetenzbegriffs von Weinert (2001a, S. 27–28):

> „Dabei versteht man unter Kompetenzen die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können."

Weinert bezieht sich damit auf im Projekt *Definition and Selection of Competencies* (DeSeCo) (Rychen & Hersh Salganik, 2001) entwickelte Bemühungen der OECD, „den vieldeutigen Leistungsbegriff generell durch das Konzept der Kompetenzen zu ersetzen" (Weinert, 2001a, S. 27–28). In der Folge seiner Veröffentlichung macht Weinert (ebenda) deutlich, dass fachliche Kompetenzen einen essentiellen Bestandteil der Erträge schulischen Unterrichts darstellen, sich aber Prioritätensetzungen zwischen diesen und *fachübergreifenden Kompetenzen* wie Problemlösen und *Handlungskompetenzen* „als höchst problematisch erwiesen" (S.28) haben. Es liegt demnach ein vielleicht durch die Auswahl des Zitats durch Klieme et al. begünstigtes Missverständnis vor, wenn man Weinert eine solche Priorisierung zuschreibt. So wie Weinert die sehr viel weiter gediehene Entwicklung bei der Messung fachlicher Kompetenzen hervorhebt, sieht er aber auch keinen Grund, diese verfügbaren Instrumente nicht zur Erkenntnisgewinnung einzusetzen. Die Untersuchung von Kompetenzstufenmodellen in der vorliegenden

[1] Die Wendung *(datenbasierte) Unterrichts- und Schulentwicklung* wird im Zusammenhang mit der neuen Steuerung immer wieder verwendet und bezieht sich auf eine Abgrenzung der Nutzung von Daten für Steuerungszwecke von Steuernden außerhalb der Schule und meint eine den Unterricht in der Schule unmittelbar betreffende Nutzung. Im Einzelfall ist ggf. zwischen Unterrichts- und Schulentwicklung zu differenzieren.

Arbeit schließt sich an diese Fokussierung an, wie sie auch in den Erhebungen zum Bildungstrend und den Vergleichsarbeiten zu finden ist, was aber keinesfalls eine besondere Bedeutung der fachlichen gegenüber überfachlichen oder handlungsbezogenen Kompetenzen reklamiert.

Kompetenzstufenmodelle sind kommunikationstechnisch hilfreiche Zerlegungen der empirisch gemessenen Kompetenzstände in Niveaustufen. Im besten Fall werden diese Zerlegungen kriterial begründet, so dass einer Schülerin oder einem Schüler mit dem Erreichen einer bestimmten Kompetenzstufe beschreibbare Fähigkeiten zugeschrieben werden können, die sich auf konkrete Anforderungen beziehen. Prenzel et al. (2013, S.40) beschreiben für PISA: „Kompetenzstufen in PISA sind demnach eine inhaltliche Interpretation der Tests und eine Form der Übersetzung von Punkten in anschauliche Kompetenzbeschreibungen der Schülerinnen und Schüler.“ Schüler*innen die der gleichen Kompetenzstufe zugeordnet werden, können Aufgaben mit etwa ähnlichen Anforderung korrekt lösen und solche, zu deren Lösung höhere Anforderungen notwendig sind, eher nicht. Im Abschnitt 3.2 werden diese Strukturen für das Verständnis der Arbeit näher erläutert[2]. Die Beschreibung dieser genauen Anforderungen rekurriert zumeist auf Kompetenzstrukturmodelle. Nach Hartig und Klieme (2006) explizieren Kompetenzstrukturmodelle die unterschiedlichen Dimensionen, die sich aus der „Struktur der interessierenden Aufgaben und Anforderungen“ ergeben. Für die Mathematik werden beispielsweise prozessbezogene von inhaltsbezogenen Kompetenzen unterschieden. Solche Strukturen bilden in den Bildungsstandards wie den Rahmenlehrplänen der Länder die Grundlage für die daran anschließend formulierten konkreten Anforderungen für einen oder verschiedene definierte Zeitpunkte der Bildungskarrieren von Schülerinnen und Schülern. Für die Feststellung von Kompetenzständen sind diese Modelle eine ausreichende Basis. Für Fragestellungen, die sich über eine Feststellung von Kompetenzständen hinaus auf die Entwicklung von Kompetenzen beziehen, sind Kompetenzentwicklungsmodelle notwendig. Denn die Entwicklung einer Kompetenz ist mit zwei verschiedenen Kompetenzständen allein nicht hinreichend beschrieben. Hierfür sind weitreichende theoretische und fachdidaktische Überlegungen anzustellen und empirisch abzusichern. Solche Modelle spielen in der vorliegenden Arbeit keine Rolle, wenngleich zu hinterfragen ist, ab wann bei einer zur Feststellung von Entwicklung mit Hilfe einer mehrfachen Verabreichung eines Instruments zur Messung von Kompetenzständen, Kompetenzentwicklungsmodelle notwendig zu hinterlegen wären.

[2]Für eine umfassendere Darstellung sei hier auf Hartig und Klieme (2006) verwiesen.

1.2. Vergleichsarbeiten und Bildungstrend

Die Vergleichsarbeiten sind eine von vier Säulen der Gesamtstrategie zum Bildungsmonitoring der Konferenz der Kultusminister*innen (KMK, 2016c), welche auf die Plöner Beschlüsse vom 02.06.2006 zurückgehen. Nachdem Regelstandards für das Ende der Primarstufe am Ende der Jahrgangsstufe 4 (KMK, 2004f, 2005f und 2005d) sowie für den Abschluss der Hauptschule am Ende der Jahrgangsstufe 9 (KMK, 2004e, 2005e, 2005c und 2005b) und den Mittleren Schulabschluss zum Ende der Jahrgangsstufe 10 (KMK, 2003c, 2004a, 2004c und 2004b) durch die KMK als Grundlage für die Rahmenlehrpläne der Länder beschlossen worden sind, definieren die vier Säulen Maßnahmen zur Überprüfung der Implementation der Bildungsstandards durch die Länder. Neben der fortgesetzten Beteiligung Deutschlands an den internationalen Assessments zur Feststellung der Leistungsfähigkeit der Bildungssysteme (wie PISA, TIMSS) und der Bildungsberichterstattung auf nationaler Ebene sowie für jedes Land, stehen im Folgenden mit den *Vergleichsarbeiten* und dem *Bildungstrend* zwei Säulen der Gesamtstrategie im Fokus.

1.2.1. Vergleichsarbeiten

Die Vergleichsarbeiten geben den Schulen ein Instrument an die Hand, mit dem diese zu zwei Zeitpunkten der Bildungsbiographie ihrer Schülerinnen und Schüler für definierte inhaltliche Domänen überprüfen können, inwieweit zentrale Kompetenzen beherrscht werden. Ergebnisberichte sind damit zentrale, kriteriale Vergleiche, sofern sie die Kompetenzstufenbezüge der Aufgaben geeignet einbeziehen. Die Zeitpunkte dieser Überprüfung wurden absichtsvoll für die Primarstufe ein und für die Sekundarstufe zwei Jahre vor den Zeitpunkt der Definition der Standards gelegt. So werden den Lehrkräfte mit den Ergebnissen der VERA-Tests in der 3. bzw. 8. Klasse Anhaltspunkte für eine Weiterarbeit gegeben, die ein Erreichen der in den Bildungsstandards beschriebenen Kompetenzen bis zum Ende der Jahrgangsstufe 4 bzw. 10 sicherstellen sollen. Auf Ihrer Webseite (IQB, 2021) ordnet das IQB mit Bezug auf Henschel und Stanat (2019) die Vergleichsarbeiten deshalb auch als *Frühwarnsystem* ein. Die Vergleichsarbeiten sind damit klar als formatives Instrument der Entwicklung von Unterricht angelegt. Da es bei dieser Einordnung aber um die Nutzung geht (Schütze et al., 2018) und nicht um das Instrument selbst, kann man auch Maier (2010) folgen und in der Implementation der Vergleichsarbeiten formative, aber überwiegend summative Aspekte finden. Über diese und weitere „funktionale Zweideutigkeiten von VERA“ berichtet auch Schliesing (2017, S. 24ff).

Unklar ist der Umgang mit dem Problem, dass sich VERA gerade nicht explizit auf die für Lehrkräfte verbindlichen landesspezifischen Rahmenlehrpläne bezieht, sondern auf die Kompetenzbeschreibungen der Bildungsstandards. Die Länder haben sich zwar dazu verpflichtet, die Gestaltung ihrer Rahmenlehrpläne an den Standards zu *orientieren*, aber die Formulierung deutet schon an, dass sie inhaltlich nicht notwendig kongruent sind, sich teilweise auch nicht des identischen Vokabulars bedienen. Die Prüfung einer angemessenen Passung der Rahmenlehrpläne steht aus, wie auch eine Festlegung, was eine solche *angemessene Passung* wäre. In Abhängigkeit von der landesspezifischen Strategie der Implementation von Bildungsstandards, muss eine Anpassung der zentralen Informationen zu den Tests sowie der Darstellung der Ergebnisse an die landeseigenen Gegebenheiten entweder vom administrierenden Landesinstitut oder von den Lehrkräften selbst geleistet werden.

Die Durchführung der Vergleichsarbeiten ist wegen der verpflichtenden Teilnahme aller Schüler*innen eines Jahrgangs an öffentlichen Schulen als Vollerhebung anzusehen. Jede Schule, jede Klasse, jede Schülerin und jeder Schüler kann eine Rückmeldung erhalten. Allerdings sind für eine gute kriteriale und/oder soziale Vergleichbarkeit Bedingungen an die Testdurchführung zu stellen, die in einem Manual für die durchführenden Lehrkräfte beschrieben sind. Zudem kann eine mögliche Inaugenscheinnahme der Tests vor der Testung oder ein gezieltes Üben der zuvor bekannten Domänen genauso wenig sicher ausgeschlossen werden, wie eine gezielte Beeinflussung der Ergebnisse bei Bewertung und Eingabe. Es konnte beobachtet werden, dass solche Einflussnahmen durch Lehrkräfte das Ergebnis (eher positiv) beeinflussen (Graf et al. (2013). Eine Klassifikation solcher und anderer nicht-intendierter Effekte findet sich bei Bellmann und Weiß (2009)).

Nicht zuletzt beziehen sich die Ergebnisse auf einen begrenzten Inhaltsbereich, operationalisiert durch abzählbar viele Aufgaben. Für die Mathematik bei VERA-8 wurden zwischen 2016 und 2020 zwei Testhefte eingesetzt, die sich aus zwei von drei Aufgabenblöcken zu je 40 Minuten zusammensetzten. Es werden demnach Mathematikaufgaben eingesetzt, die insgesamt eine Bearbeitungszeit von 120 Minuten beanspruchen, eine Testperson muss dabei ca. 80 Minuten Bearbeitungszeit aufbringen. Die Aufgaben selbst sind ein wesentlicher Bestandteil der Rückmeldungen. Den Schulen werden nicht nur Testhefte für Schüler*innen und Manuale für Lehrkräfte geliefert, sondern auch Materialien, die sich teilweise sehr konkret mit den Aufgaben auseinandersetzen, ihren Gehalt für die diagnostische Analyse in besonderer Weise herauszustellen versuchen, auch, um so eine fachdidaktisch intendierte, kompetenzorientierte Weiterarbeit mit den Ergebnissen anzuregen. Teilweise erfordert die Form solcher

Nutzung, dass Gruppen, deren Auswertungen verglichen werden sollen, mit identischen Aufgaben getestet werden. Auch weil den Lehrkräften mit dem gelieferten Materialien die den Bildungsstandards zugrundeliegende Idee kompetenzorientierten Unterrichts an immer neuen Aufgaben näher gebracht werden soll, wird im Allgemeinen jedes Jahr ein Set neuer, unbekannter Aufgaben zur Verfügung gestellt. Die jährlichen Tests eröffnen allerdings auch die Möglichkeit einer Nutzung von Ergebnissen als Monitoring auf unterschiedlichen Ebenen.

Wie die Zusammenfassung von Tarkian et al., 2019 deutlich zeigt, ist die Nutzung der Ergebnisse aus den Vergleichsarbeiten nicht allein auf Aspekte der Schul- und Unterrichtsentwicklung beschränkt. Die Beschlüsse der KMK nennen dies zwar dezidiert als Kernziel von VERA (KMK, 2018b), schließen Anderes aber auch nicht aus. So werden VERA-Ergebnisse teilweise für Gespräche zwischen Schulaufsichten und Schulleitungen oder im Rahmen von Schulinspektionen genutzt. In Berlin fließen aktuell aggregierte Ergebnisse jeder Schule in ein die Schulen beschreibendes Indikatorensystem (Senatsverwaltung für Bildung, Jugend und Familie, 2021) ein[3], im Land Brandenburg werden die Ergebnisse im Zentralen System zur Online-Verwaltung von Schulinformationen (ZENSOS) (Ministerium für Bildung, Jugend und Sport, 2021) vorgehalten. Für die in dieser Arbeit betrachteten Aspekte der Nutzung, der Validität und der Abrufe von Ergebnissen, sind offizielle Verlautbarungen zur Zielstellung von VERA von herausgehobener Bedeutung. Im Abschnitt 1.3 wird dem deshalb mit einer Quellenanalyse nachgegangen.

1.2.2. Bildungstrend

Das IQB ist neben der Entwicklung der Aufgaben und Materialien für Vergleichsarbeiten unter anderem mit der Aufgabenentwicklung, aber auch der Durchführung der Bildungstrends befasst. Auch diese überprüfen das Erreichen der Bildungsstandards von Schülerinnen und Schülern. Getestet wird dabei in der Jahrgangsstufe 4, also genau an dem Zeitpunkt, für den die Standards für die Primarstufe beschrieben wurden sowie in der Jahrgangsstufe 9. Für das Ende der Sekundarstufe I liegen Bildungsstandards für verschiedene Abschlüsse vor. Die Überprüfungen finden ein Jahr nach den Vergleichsarbeiten statt. Während ein KMK-Beschluss den Vergleich von VERA-Ergebnisse über Länder hinweg ausschließt (KMK, 2012a, S.3), stehen beim Bildungstrend die Leistungen der Schülerinnen und Schüler aller 16 Länder im Zentrum der Analyse, weniger aber im Sinne eines Rankings, was mit der neuen

[3]Die Nutzung von Ergebnissen aus den Vergleichsarbeiten bezog sozial adjustierte Landesergebnisse ein. Pandemie-bedingt fand VERA-3 im Jahr 2020 nicht statt und 2021 wurden VERA-3 sowie 8 nur als freiwillige Tests durchgeführt, so dass keine verlässlichen Äquivalente berechnet werden konnten. Deshalb blieb der Einbezug von VERA in diesen Jahren aus.

Titelsetzung *Bildungstrend* statt *Ländervergleich* unterstrichen werden soll. Der Bildungstrend ist als summatives Instrument des Bildungsmonitorings angelegt und informiert primär die Bildungsadministrationen aber auch die Öffentlichkeit über den Leistungsstand der Schüler*innen und damit über das Gelingen der Implementation der Bildungsstandards.

Für die Einschätzung des Leistungsstandes reicht die Erhebung einer Stichprobe von Schülerinnen und Schülern aus jedem Land aus. Anders als bei VERA, wo die Lehrkräfte der Schulen selbst die Tests administrieren, finden die Testungen im Rahmen des Bildungstrends testleiter*innenbasiert statt. Auch die Korrektur der Tests, bei VERA im Allgemeinen Aufgabe der den Test durchführenden Lehrkräfte, wird hier von geschulten Kodiererinnen und Kodierern übernommen. Durch den vergleichsweise großen Aufwand beim Bildungstrend soll eine hohe Durchführungsobjektivität gewährleistet werden.

Dem Ziel folgend, auf Landesebene stabile Leistungsparameter zu ermitteln, die über mehrere Testzyklen Trendvergleiche ermöglichen, werden beim Bildungstrend bei jeder Testung überwiegend identische Aufgaben verwendet, die zuerst zu Aufgabenblöcken zusammengefasst auf verschieden Testhefte verteilt werden (Becker et al., 2019). Für die Testung der Mathematik im Jahr 2012 in der Sekundarstufe wurden beispielsweise 31 Aufgabenblöcke mit je 20 Minuten Testzeit entwickelt, so dass Aufgaben mit einer Testzeit von insgesamt mehr als 10 Stunden eingesetzt wurden (Siegele et al., 2013). Die Testhefte mit jeweils 6 Aufgabenblöcken werden zufällig auf die Schülerinnen und Schüler verteilt. Mit diesem sogenannten Multi-Matrix-Design wird eine effiziente Testung bei umfassender Abdeckung der Inhaltsbereiche gewährleistet.

Die Vergleichsarbeiten und der Bildungstrend, zwei Erhebungen, die auf Beschlüsse der Kultusministerkonferenz (KMK) zum Bildungsmonitoring zurückgehen, weisen damit erhebliche Differenzen in der Konstruktion ihrer Instrumente und der Durchführungen auf, die sich aus den deutlich unterschiedlichen Zielstellungen ableiten lassen (Tabelle 1.1).

1.3. Zielstellung der Vergleichsarbeiten

Als Fortschreibung der Darstellung bei Wacker und Kramer (2012, S.688) sind die in den offiziellen Dokumenten der KMK benannten Zielstellung für die Mikroebene des Unterrichts, die Mesoebene der Schule und für die Makroebene des Systems differenziert in der Tabelle A.1 im Anhang chronologisch zusammengefasst. Zweifelsohne liegt das Gewicht der Ziele auf Prozessen der Qualitätsentwicklung und -sicherung im Rahmen datengestützter Unterrichts- und Schulentwicklung.

Tabelle 1.1.: Unterschiede und Kongruenzen zwischen Bildungstrend und Vergleichsarbeiten

	Bildungstrend	Vergleichsarbeiten
Testentwicklung	IQB	IQB
Testdurchführung	IQB (geschulte Testleiter*innen)	Länder (Lehrkräfte selbst)
Korrektur der Tests	IQB (geschulte Kodierer*innen)	Länder (i.A. Lehrkräfte selbst)
Testzeitpunkte		
Primarstufe	Klasse 4	Klasse 3
Sekundarstufe	Klasse 9	Klasse 8
Testgruppe	Stichprobe	Vollerhebung (öffentliche Schulen)
Design	Multi-Matrix	2 Testversionen nach Schulform
Zielstellung	Systemmonitoring	indifferent (Unterstützung von Schul- und Unterrichtsentwicklung, Monitoring)
Testinhalte		
Basis	Bildungsstandards	Bildungsstandards
Abdeckung der Inhalte	mehr als 10h	ca. 2h
Rückmeldung	Verteilung auf Kompetenzstufen für Länder und Schulformen, Untergruppen von Schüler*innen auf Landesebene	Verteilung auf Kompetenzstufen sowie Lösungshäufigkeiten Einzelaufgaben und Aufgabengruppen, für Untergruppen von Schüler*innen bis Einzelschüler

Dass noch vor dem alle Länder einschließenden Start der Vergleichsarbeiten Ziele formuliert wurden, die Analysen und ggf. Maßnahmen auch auf Ebene des föderalen Systems betreffen, ist auch einer hier noch fehlenden Abgrenzung von den späteren Ländervergleichen zuzuschreiben. Eine tatsächliche Vermischung von Zielen findet sich auf der Mesoebene. Diese schulbezogenen Analysen betreffen einerseits Prozesse schulintern organisierter Qualitätsentwicklung, andererseits aber auch solche, bei denen schul- oder auch klassenbezogene Daten mit der Schulaufsicht bzw. der Schulinspektion Stakeholdern außerhalb der Schule zugänglich werden. Die zentral formulierten Zielstellungen betreffen in erster Linie die Arbeit der Lehrkräfte auf der Mikroebene, die Planung und Entwicklung von Unterricht und auch unmittelbar die Förderung der Schüler*innen als Maßnahme, die sich als Schlussfolgerung aus

der Analyse der Daten der Vergleichsarbeiten ergeben.

Es gibt aber eine zweite Zielstellung, die nicht gleich zu Beginn, aber in der letzten Dekade konsequent parallel neben die Unterrichts- und Schulentwicklung gestellt wird: die Vermittlungsfunktion. Mit ihrer Einführung verpflichteten sich die Länder auf die Implementation der fachlichen und fachdidaktischen Konzepte der Bildungsstandards.

> „Mit der Einführung von Bildungsstandards wollen die Länder für Transparenz hinsichtlich der schulischen Anforderungen sorgen, die Entwicklung eines an Kompetenzen orientierten Unterrichts unterstützen, die gezielte Förderung von Schülerinnen und Schülern verstärken und Rechenschaft über erreichte Ergebnisse ablegen.“ (KMK, 2006b, S.11)

Dies wird durch die Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung (KMK, 2010) konkretisiert und in der Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (KMK, 2012b) klar als Vermittlungsfunktion benannt. Den Ergebnisrückmeldungen kommt dabei als Rekonstruktion der Konzepte für die eigene Klasse eine herausgehobene Bedeutung zu.

Für die Diskussion in dieser Arbeit wird die Zusammenfassung der im Anhang A.1 dargelegten Quellenanalyse zur Zielbestimmung der Vergleichsarbeiten im Folgenden auf zwei Aspekte fokussiert, die schon in den von der KMK beauftragten Expertisen von Weinert (2001b) und Klieme et al. (2003) aufgeworfen worden sind: die *Ownership of Data* und die Bedeutung von *individuellen Analysen*. Weinert (2001c) betont überdies, dass die Verwertung von Ergebnissen ein Prozess sein muss, in dem Wissenschaft und Praxis notwendig aufeinander angewiesen sind. Bei Klieme et al. (2003, S.82-83) wie bei Weinert (2001c) werden die unterschiedlichen Studienformen - insbesondere das Systemmonitoring und die Individualdiagnostik - voneinander abgegrenzt und auf die unterschiedlichen Verwertungslogiken eingegangen.

Ownership of Data

In der Pressemitteilung vom Mai 2002 (KMK, 2002b) werden zur Beschreibung der Zielstellung der Vergleichsarbeiten das erste Mal zwei Aspekte nebeneinandergestellt:

- die Überprüfung des Erreichens der Bildungsstandards durch die Länder, also ein Monitoring und

- die Ermöglichung einer optimalen Förderung von Schülerinnen und Schüler in der Schule als Ergebnis von Unterrichts- und Schulentwicklung.

Dabei sind die Ergebnisrückmeldungen für Lehrkräfte sowohl in allen KMK-Dokumenten wie auch in den Veröffentlichungen der Länder das zentrale Anliegen und die Arbeit der Lehrkräfte wird in der Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (KMK, 2012b) als *Kernfunktion des Instruments* bezeichnet. Diese Funktion wird durch die oft wiederholte Aussage unterstrichen, dass mit der Lieferung von immer neuen Aufgaben zur Verwendung durch die Lehrkräfte auch nach dem Test, die Implementation der Bildungsstandards unterstützt werden soll. Des Weiteren unterscheiden sich die landesspezifischen VERA-Implementationen aber bezüglich der Nutzung von Ergebnissen im Sinne eines Monitorings. Unbedeutender scheint hier zu sein, ob alle Lehrkräfte, Fachkonferenzleitungen und die Schulleitung auf alle schulinternen Daten zugreifen können, oder jeweils nur auf einen für sie bestimmten Ausschnitt. Als Teil der Ownership of Data ist die zentrale Frage, ob schulinterne Daten eine außerschulische Rezeption und Nutzung erfahren.

Erhält die Schulaufsicht automatisch einen Zugriff auf die Ergebnisse? Muss die Schule diese Daten der Inspektion zuliefern bzw. hat die Inspektion direkten Zugriff? Können die Daten durch die Administration eingesehen werden oder werden sie durch diese für bestimmte Zwecke genutzt, ggf. zur Zuteilung von Ressourcen? Oder werden Ergebnisse gar zentral veröffentlicht? Die Festlegungen in den verschiedenen KMK-Papieren dazu sind selten ausschließend: Eine landesinterne Nutzung von VERA-Daten ist *möglich*, es liegt *in der Entscheidung der Länder*, inwieweit die Schulaufsicht die Schule bei der Datennutzung unterstützt und berät, dabei *kann* die Nutzung der VERA-Daten verbindlicher Gegenstand von Bilanzierungsgesprächen werden. Und letztendlich soll von einer Veröffentlichung von VERA-Ergebnissen *abgesehen werden*, ausgeschlossen wird dies aber nicht.

Die Projektbeschreibung der ersten Implementation durch Helmke und Hosenfeld (2003a) an der Universität Koblenz-Landau schließt eine Rückmeldung schulbezogener Daten an hierarchisch übergeordnete Stellen außerhalb der Schule dezidiert aus, erlaubt aber die landesbezogene Auswertung einer (anonymen) Stichprobe für die Administration. Das Erkennt gleichermaßen die herausgehobene Qualität der Daten einer zentralen Erhebung von standardisierten Testinhalten für die Administration und deren Interesse an solch einer Statusfeststellung an, aber eben gleichermaßen bzw. zuvorderst die Eigenverantwortung und Zuständigkeit der Lehrkräfte und der Beteiligten innerhalb der Schulen. Allerdings richtete sich diese Standortbestimmung von Helmke und Hosenfeld an die Administration nur ihres Landes.

Mit Bezug auf Schildkamp und Teddlie (2008) fassen Maier und Schymala (2011) zusammen, dass „Lehrkräfte nur dann sinnvoll mit zentralen Testrückmeldungen umgehen, wenn sie diese als ihre eigenen Evaluationsdaten akzeptieren“ und sie schließen damit an die Weinert'sche Forderung der Kooperation bei der Ausgestaltung der Ergebnisrückmeldung zwischen Wissenschaft und Praxis an. Das Instrument, der fachinhaltliche Fokus, der Zeitpunkt und das Reglement der Durchführung sind in hohem Maße von außen bestimmt. Für die Lehrkräfte bleibt bei den Vergleichsarbeiten allein beim Umgang mit den Ergebnissen „die Möglichkeit der Aneignung des Externen“ (Pant, 2011).

Eine Stärkung der innerschulischen Nutzung erfährt das Instrument in Zukunft sicher durch mehr Gestaltungsraum für Lehrkräfte bei der Auswahl von Testheften mit inhaltlichen oder schwierigkeitsbezogenen Schwerpunkten (Modularisierung). Eine konsequente Ablehnung der schulexternen Nutzung würde diese Stärkung vermutlich weiter befördern. Klar ist aber auch, dass eine stärkere Modularisierung einer Verwertung der Ergebnisse im Sinne eines Monitorings entgegensteht, mindestens aber in besonderer Weise statistischer Absicherung bedarf.

So uneindeutig die Aussagen zur Nutzung von Ergebnissen aus den Vergleichsarbeiten für administrative Zwecke in den Dokumenten der KMK sind, verwundert es kaum, dass sich in den Ländern ein großes Spektrum von Interpretationen findet, angefangen von Thüringen und Hessen, wo selbst anonyme wissenschaftliche Auswertungen der Zustimmung jeder Einzelschule bedürfen[4], bis hin zu Ländern, wo der Einbezug von VERA-Ergebnissen in Qualitätsprofile eine übliche Praxis darstellt (siehe Tarkian et al., 2019, S.54ff).

Individuelle Analyseebene

Schon 2003 wird bei Helmke und Hosenfeld (2003a) aus methodischen und inhaltlichen Gründen auf einen lediglich ergänzenden Charakter von VERA-Ergebnissen auf Ebene von Schüler*innen für die Schulpraxis verwiesen. Dies und den erheblichen Messfehler brachten auch Klieme et al. (2003) gegen eine Auswertung auf dieser Ebene vor. Parallel dazu verzichtet die erste Gesamtstrategie der KMK (KMK, 2006c) auf die Nennung der Analyseebene *Schüler*innen*. Untersucht wird der Leistungsstand der *Schulen* und *Klassen*, das Ergebnis soll für die Förderung der *Klassen* genutzt werden. In der *Konzeption zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung* (KMK, 2010) gute 3 Jahre später, zeigt die Verteilung der Leistungen auf Niveaustufen immerhin „in sehr eingeschränktem Maße“ auch für Einzelschüler*innen potentiellen Förderbedarf an. Mit der ersten Durchführung von

[4]Für diese Festlegung konnte in keinem der beiden Länder eine (ggf. schulgesetzliche) Regelung gefunden werden, so dass davon auszugehen ist, dass sie lediglich einer gelebten Praxis entspricht.

VERA mit im IQB entwickelten Aufgaben 2010 und der damit verbundenen Anbindung an die Metrik der Bildungsstandards, wird bei der Bewertung von Ergebnissen für Einzelschüler*innen auf den Zusammenhang mit anderen Informationen verwiesen. Etwas konkreter wird zwei Jahre später (KMK, 2012b) ausgeführt, dass eine individuelle Ergebnisrückmeldung für Schüler*innen fachlich vertretbar sei, wenn diese pädagogisch angemessen eingeordnet würde. Dass diese nicht obligatorisch erfolgen muss, wie das Dokument feststellt, haben die Länder anders geregelt. In allen 16 Ländern gehören individuelle Ergebnisrückmeldungen zum Standardportfolio (siehe auch Tarkian et al., 2019, S. 65ff). Mit der Modernisierung der Vergleichsarbeiten (KMK, 2018a) werden diese nun als pädagogisches Diagnoseinstrument bezeichnet und damit, dass die Auswahl von Ergänzungsmodulen so erfolgen soll, dass sie dem Kompetenzniveau der Schüler*innen besonders gut entspricht, scheinen sich bezüglich der Auswertung der Vergleichsarbeiten auf individueller Ebene schulpraktische Bedarfe ein Stückweit gegen psychometrische Bedenken durchgesetzt zu haben. Aus Sicht des Autors fehlt hier eine zwischen Wissenschaft und Schulpraxis angemessen ausgelotete Position.

1.4. Genese und Entwicklung der Vergleichsarbeiten

Die oft als PISA-Schock bezeichnete Situation nach der Auswertung der Ergebnisse der PISA-Studie aus dem Jahr 2000, in der Deutschland kaum das Mittelfeld der beteiligten Länder erreichte, führte wie in vielen Ländern auch in Deutschland zu öffentlicher Resonanz und in deren Folge zu gravierenden Veränderungen. Eine Folge war der Auftrag zur Erstellung einer Expertise (Klieme et al., 2003), die als Start- und Orientierungspunkt für die Entwicklung bundesweiter Bildungsstandards angesehen werden muss. Ab 2004 wurden im Auftrag der KMK Regelstandards für drei explizite Zeitpunkte der Bildungslaufbahn beschrieben:

- für das Ende der vierten Jahrgangsstufe, welches in der übergroßen Mehrzahl der Länder das Ende der Primarstufe markiert, und zwar für die Fächer Deutsch und Mathematik.
- für das Ende der Sekundarstufe I, also dem Zeitpunkt der Prüfungen zum Mittleren Schulabschluss bzw. zum Hauptschulabschluss (und landesspezifischer Äquivalente), zusätzlich zu Deutsch und Mathematik auch für die erste Fremdsprache (beschränkt auf Englisch und Französisch). Später folgten auch Standards für die Naturwissenschaften.
- für das Abitur für alle bisher benannten Fächer.

Die Vergleichsarbeiten starteten als politisches Projekt der sozialliberalen Koalition nach der Wahl 2001 in Rheinland-Pfalz. Im Koalitionsvertrag wurde die Einführung von Vergleichs-

arbeiten in der 4. Klassenstufe in den Fächern Deutsch und Mathematik verabredet und mit dem Beschluss des Landtages vom 25.04.2002 eingeleitet (Landtag Rheinland-Pfalz, 23. Wahlperiode, 2002). Anlass für diesen Schritt waren die öffentlichen Diskussionen um die Ergebnisse Deutschlands bei den internationalen Bildungsstudien TIMSS und vor allem PISA (Lorenz, 2005). Die Zielstellung war schon damals unscharf; benannt wurde sowohl ein Systemmonitoring wie auch die Unterstützung der Selbstevaluation von Lehrkräften. So formulierte Doris Ahnen, damalige Bildungsministerin in Rheinland-Pfalz, mehrere Ziele, wie die individuelle Förderung der Kinder, das Anregen von Reflexionen bei Lehrkräften, insbesondere über die eigene Diagnosefähigkeit, den Vergleich zwischen Klassen, auch zwischen Schulen und ebenso sollten die Vergleichsarbeiten Anstoß für pädagogische und fachdidaktische Diskussionen geben (Bildungsklick, 2005). Die gleiche Pressemeldung interpretiert aber auch Landesergebnisse der ersten flächendeckenden Testung für Deutsch und Mathematik. Diese Vermischung von zwei so unterschiedlichen Zielen wird die Vergleichsarbeiten über die Jahre genauso begleiten wie die Diskussionen darum. Für ein Systemmonitoring der Länder, wie dies damals schon durch die Erweiterung von PISA mit PISA-E erfolgte und später durch die Ländervergleiche (den heutigen Bildungstrends) durch das IQB fortgesetzt wurde, reicht die Untersuchung einer Stichprobe jedes Landes aus (vergleiche Abschnitt 1.1). Um aber eine obligatorische Evaluation des Unterrichts sicherzustellen und mit der Idee der Unterrichtsentwicklung zu verbinden, wurden die Vergleichsarbeiten flächendeckend umgesetzt.

Mit Konzeption, Vorbereitung und letztlich auch der Durchführung der ersten Vergleichsarbeiten für Rheinland-Pfalz wurde das Institut für Psychologie der Universität Koblenz-Landau unter der Verantwortung von Prof. Helmke und Prof. Hosenfeld betraut. 2003 wurden die Vergleichsarbeiten das erste Mal lediglich im Fach Mathematik durchgeführt. Schon im Herbst fand in Berlin als Vorlauf für eine Beteiligung ab 2004 eine Erprobung statt. Und auch alle weiteren fünf zu diesem Zeitpunkt SPD-regierten Länder Brandenburg, Bremen, Mecklenburg-Vorpommern, Nordrhein-Westfalen und Schleswig-Holstein schlossen sich dem Vorhaben an, um „Qualitätsentwicklung und Qualitätssicherung in der Grundschule als Aufgabe über Ländergrenzen hinweg gemeinsam zu gestalten.“(Senatsverwaltung für Bildung, Jugend und Sport, 2003).

Die Vergleichsarbeiten wurden durch die Universität Koblenz-Landau als Test in der Klassenstufe 4 entwickelt und 2008 auf die Klassenstufe 3 vorgezogen, um die Ergebnisse davor zu schützen, für die Übergangsentscheidung in die weiterführende Schulen verwendet zu werden. In der Mehrzahl der Länder findet dieser Übergang nach der Klassenstufe 4 statt. Parallel

dazu wurde von der KMK das IQB als Institut aller Länder gegründet (KMK, 2004d) und unter Anderem damit beauftragt, Tests zu entwickeln, die es der einzelnen Lehrkraft einer Klasse ermöglichen festzustellen, inwieweit ihre Schülerinnen und Schüler die in den Standards ausgewiesenen Kompetenzen erreichen (KMK, 2006b). 2008 wurde der erste Test im Rahmen der Vergleichsarbeiten in der Jahrgangsstufe 8 vom IQB durchgeführt. Nach einer bundesweiten Neuausschreibung ging die Entwicklung der VERA-3 ebenso an das IQB über. Das IQB liefert seitdem für VERA-3 und VERA-8 jährlich Testhefte mit didaktischen Kommentierungen, Durchführungsanleitungen inkl. der Hinweise zur Korrektur und Bewertung der Aufgaben sowie ein technisches Manual, welches über den Prozess der Testentwicklung informiert und Hinweise zur Psychometrie der Test zur Verfügung stellt. Das IQB ist zudem mit der Durchführung des Bildungstrends wie der Entwicklung der Kompetenz- und Kompetenzstufenmodelle betraut (zuerst KMK, 2006b; und später KMK, 2016a). Als Konsequenz der Verlinkung der Testinstrumente werden die Ergebnisse aus den VERA-Tests auf der gleichen Metrik wie der des Bildungstrends berichtet, können also die Fähigkeiten jeder Schülerin und jedes Schülers auf die Kompetenzstufen der KMK bezogen werden. Für die Vergleichsarbeiten wird bei einer Ergebnisrückmeldung auf dieser Ebene allerdings auf das große Vertrauensintervall hingewiesen (zum Beispiel Leutner et al. (2008) S.151 ff, sowie weiter unten im Abschnitt 3).

Die Implementationen der Vergleichsarbeiten in den Ländern verliefen unterschiedlich. Die genauen Festlegungen zur Teilnahme haben sich über die Jahre mehr und mehr ausdifferenziert. Ein umfassender Überblick findet sich bei Tarkian et al. (2019). Allerdings sind die darin festgehaltenen Durchführungsmodalitäten der 16 Länder schon zum Erscheinungstermin nicht mehr vollständig zutreffend (zum Beispiel Niedersächsisches Kultusministerium, 2019; MBJS, 2020). Die für das Ressort Bildung zuständige Senatsverwaltung des Landes Berlin[5] und das Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (MBJS) gründeten 2006 das Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ), welches seitdem für beide Länder unter Anderem mit der Durchführung der Vergleichsarbeiten betraut ist.

Für die Jahrgangsstufe 3 liegen jährlich neue Testunterlagen für die Fächer Deutsch und Mathematik vor. Dabei wird für die Domäne *Deutsch Lesen* jedes Jahr ein Testheft bereitgestellt und durch ein zweites Heft für eine der drei Domänen *Rechtschreiben*, *Zuhören* und *Sprache und Sprachgebrauch untersuchen* ergänzt, wobei die zwei Tests an zwei unterschiedli-

[5] Zum Zeitpunkt der Gründung des ISQ 2006 umfasste die Bildungsverwaltung die Ressorts *Bildung, Jugend und Sport*, in der folgenden Legislaturperiode von Ende 2006 bis Ende 2011 *Bildung, Wissenschaft und Forschung* und danach bis heute *Bildung, Jugend und Wissenschaft*.

chen Tagen den Schüler*innen vorgelegt werden sollen. In Mathematik werden in jedem Jahr zwei Leitideen fokussiert, die in einem Testheft nacheinander angeordnet sind. Dabei wiederholt sich jedes Jahr eine der zwei Leitideen aus dem Vorjahr und wird um eine Neue ergänzt. So wird jede der fünf Leitideen *Zahlen und Operationen*, *Muster und Strukturen*, *Größen und Messen*, *Raum und Form* sowie *Daten, Häufigkeiten und Wahrscheinlichkeit* wechselnd immer zwei Jahre nacheinander getestet und nach fünf Jahren startet ein neuer Turnus. Alle Tests bei VERA 3 nehmen zwischen 30 und 40 Minuten in Anspruch. Eine Übersicht über die vom IQB (und bis 2009 von der Universität Koblenz-Landau) bereitgestellten Tests und die Verbindlichkeit der Durchführung im Land Berlin finden sich in der Tabelle A.2 im Anhang A.3.

Auch in der Jahrgangsstufe 8 wird die Domäne *Deutsch Lesen* in jedem Jahr getestet und durch eine der drei anderen Domänen ergänzt. Im Jahr 2011 wurde durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) nach Abstimmung in der zuständigen Steuergruppe die Domäne *Schreiben* als VERA-Testteil vorbereitet, dann aber nur in wenigen Ländern und oft nur freiwillig eingesetzt. Die praktische Umsetzung der Auswertung, so wurde befürchtet, bereite den Lehrkräften deutlich mehr Aufwand als bei anderen Testdomänen, so dass eine verlässliche Bewertung durch die Lehrkräfte bezweifelt wurde. Ähnlich problematisch wurden auch die Tests zur Schreibkompetenz für Englisch und Französisch als erste Fremdsprache bewertet, weshalb auch diese nur für 2010 entwickelt und auch da nur partiell eingesetzt wurden. Ansonsten wurde für beide Fremdsprachen jedes Jahr ein Test zum Lese- und Hörverstehen entwickelt. Der Mathematik-Test umfasst, anders als für die Primarstufe, jedes Jahr Aufgaben aus allen fünf Leitideen der Bildungsstandards *Zahl*, *Messen*, *Raum und Form*, *Funktionaler Zusammenhang* sowie *Daten und Zufall*. Für jedes Fach wurden zwischen 80 und 90 Minuten Testzeit beansprucht. Eine Übersicht über die vom IQB entwickelten Tests, der Verteilung auf die Schulformen sowie die Verbindlichkeit der Durchführung im Land Berlin sind der Tabelle A.3 im Anhang A.3 zu entnehmen.

Als die Universität Koblenz-Landau 2004 das erste Mal VERA in Mathematik und Deutsch durchführte, waren die Bildungsstandards noch in Entwicklung. Trotzdem wurden in den Rückmeldungen schon verschiedene Grade der Erreichung von Kompetenzen als Fähigkeitsniveaus ausgewiesen (zitiert aus Helmke & Hosenfeld, 2007b; vergleiche auch Helmke & Hosenfeld, 2007a).

- *nicht auswertbare Leistung*: Liegen keine oder unvollständige Daten vor, ist unter Umständen eine Zuordnung zu den beschriebenen Fähigkeitsniveaus nicht möglich.

- *Fähigkeitsniveau 1*: Grundlegende Fähigkeiten, einfache Aufgaben mit grundlegenden Anforderungen werden hinreichend sicher gelöst [6].

- *Fähigkeitsniveau 2*: Erweiterte Fähigkeiten, Aufgaben mittleren Anforderungsniveaus werden hinreichend sicher gelöst.

- *Fähigkeitsniveau 3*: Fortgeschrittene Fähigkeiten, es werden auch anspruchsvollere Aufgaben hinreichend sicher gelöst.

Die Entwicklung der Items, deren Normierung, die inhaltliche Beschreibung der Fähigkeitsstufen und der Einsatz der Tests in der Fläche führten zu einer stetigen Weiterentwicklung der Stufenbeschreibungen, unter Einbezug der Expertise von Fachdidaktiker*innen, Fachwissenschaftler*innen, Psycholog*innen und praktisch tätigen Lehrerinnen und Lehrern.

Ab 2010 oblag die Erarbeitung der Tests dem IQB. Die Vergleichsarbeiten in der 8. Klassenstufe wurden 2008 das erste Mal in Mathematik und ab 2009 auch in Deutsch und der ersten Fremdsprache vom IQB entwickelt. Die in umfangreichen Normierungen im Rahmen des Bildungstrends fundierte Metrik der Bildungsstandards wurde durch ein Standard-Setting (siehe auch Abschnitt 3) in inhaltlich beschriebene Kompetenzstufen unterteilt.

- *Mindeststandards nicht erreicht*

- *Mindeststandard*

- *Regelstandard*: von der KMK beschlossene Standards

- *Regelstandard plus*

- *Optimalstandard* (zuerst Maximalstandard)

Bei der Entwicklung der Vergleichsarbeiten im IQB wurden eben diese Kompetenzstufenmodelle zugrunde gelegt, die auch für den Bildungstrend eine kriteriale Einordnung der Ergebnisse erlauben.

Die Länder Berlin und Brandenburg haben beginnend mit der ersten Testung die Rückmeldung mit Bezug auf diese Kompetenzstufen in den Mittelpunkt der Berichterstattung gestellt (Emmrich & Harych, 2009).

[6] Der Passus *hinreichend sicher* verweist hier auf eine Lösungshäufigkeit von 62,5%.

Tabelle 1.2.: Kombination von 4 Modulen zu 3 Testheftversionen unterschiedlicher Schwierigkeit.

Modul 1 einfache Aufgaben	Modul 2 einfache bis mittelschwere Augaben	Modul 3 mittelschwere bis schwere Augaben	Modul 4 schwere Aufgaben
Testheft 1 für Hauptschüler*innen			
	Testheft 2 für Realschüler*innen		
		Testheft 3 für Gymnasiast*innen	

1.5. Testheft mit angemessener Schwierigkeit

Für die Sekundarstufe I entwickelte das IQB von Beginn an unterschiedliche Testheftversionen. Sie sollten auf das Leistungsniveau der Schülerinnen und Schüler angepasst sein und berücksichtigen, dass die KMK Standards zwischen Haupt- und Realschulen differenziert. Die einfachste *Version A* des Tests fokussierte auf Schüler*innen, bei denen davon auszugehen ist, dass sie einen Hauptschulabschluss bzw. einen, für das jeweilige Bundesland äquivalenten Abschluss anstreben. Das mittlere Testheft, *Version B*, richtete sich an solche Schülerinnen und Schüler, die einen Realschulabschluss anstreben, also aller Voraussicht nach den Mittleren Schulabschluss am Ende der 10. Klasse erlangen werden. Mit der *Version C* orientierte sich das schwierigste Testheft an der Leistungsfähigkeit von Schüler*innen an Gymnasien. Dazu entwickelte das IQB für jede Domäne vier Testmodule, die zu den drei Testheften kombiniert wurden (siehe Tabelle 1.2).

Der Einsatz der unterschiedlichen Testheftversionen variierte sowohl von Land zu Land wie auch über die Jahre. Ziel war es, allen Schüler*innen ein Testheft mit angemessener Schwierigkeit zur Verfügung zustellen. Aus testtheoretischer Perspektive kann gezeigt werden, dass dies bei einer mittleren Lösungshäufigkeit von 50% gegeben ist. Entsprechend der vom IQB gewählten Zuschreibung geschah eine Zuteilung häufig auf der Basis der Schulformen, auch wenn deutlich wurde und an Hand von Prüfungsstatistiken schon vorher deutlich war, dass diese sich für ein Land in großen Bereichen der Leistungsverteilung überschnitten, also keinesfalls eine eindeutige Zuweisung nach Leistungsfähigkeit darstellte. Zudem zeigten sich die zum Beispiel in den Erhebungen zum Bildungstrend bekannten Leistungsdifferenzen zwischen den Ländern. So passt die Testheftversion A bezogen auf die mittlere Lösungshäufigkeit vielleicht genauso gut zur Hauptschule in Bayern wie zur Integrierten Sekundarschule in Berlin. Aus motivationspsychologischer Sicht ist eine durchschnittliche Lösungshäufigkeit von 50% keine optimale Wahl. Das Setting der VERA-Tests erinnert Schülerinnen und Schüler an eine

Klausur, bei der man üblich mit 50% gerade noch bestanden hat. Unabhängig davon, dass es deshalb eine Erläuterung gegenüber den Schüler*innen wie den Eltern geben sollte, wurde durch die VERA-Steuergruppe im Lasten-Pflichtenheft des IQB zur Lieferung der VERA-Aufgaben eine für die Schulformen angepasste durchschnittliche Lösungshäufigkeit von 50 bis 60% angegeben. Diese Angabe bezog sich auf die Stichprobe, die durch das IQB für die Pilotierung/Normierung der entwickelten Items bundesweit verwendet wurde. Zwischen den Ländern variieren die tatsächlichen Lösungshäufigkeiten selbstverständlich.

An der Abbildung 4.5 auf Seite 85 für das Fach Mathematik bzw. den weiteren Abbildungen A.1 und A.2 im Anhang A.4 für die Domänen Deutsch Lesen (S.265) und Englisch Leseverstehen (S.267) erkennt man den Versuch, eine optimale Schwierigkeitsabstufung durch die Heftversionen herzustellen. Im Zuge des Übergangs mehrerer Länder von einem drei- zu einem zweigliedrigen Schulsystem, wurde es für das IQB zunehmend schwieriger, für die Pilotierung in jedem Land eine geeignete Stichprobe zu rekrutieren. Haupt- und Realschüler*innen waren dann im Allgemeinen in einer Schulform, oft in einer Klasse gemischt. In Abstimmung mit den Ländern wurde ab dem Jahr 2016 neben dem gymnasialen nur noch eine weitere Testheftversion zur Verfügung gestellt. Die mittleren Schwierigkeiten der Testhefte in den Abbildungen zeigen deutlich, dass dabei lediglich aus den zwei einfacheren Testheften eine neue Version entstand, deren Lösungshäufigkeit zwischen den zwei Versionen liegt, das schwierige Testheft aber weiterhin das gymnasiale Niveau abbildet.

Während die Bildungsstandards für das Ende der Sekundarstufe I zumindest zwischen dem Hauptschul- und dem mittleren Schulabschluss differenzierten, was in der Beschreibungen von integrierten Kompetenzstufenmodellen seinen Niederschlag fand, gab es für alle Schülerinnen und Schüler der Primarstufe identische Standards und in der Konsequenz auch ein einziges Testheft mittlerer Schwierigkeit für jede Domäne. Bestehende Leistungsdifferenzen zwischen den Ländern konnten hier nicht ausgeglichen werden und so war das Justieren der mittleren Schwierigkeit und die Diskussion mit den Ländern oft Thema der VERA-Steuergruppe. Öffentlich sichtbar wurde dies, als die Senatsbildungsverwaltung Berlins als Reaktion auf massive Proteste wegen zu schwerer VERA-3-Tests ankündigte, dass die Testhefte „im unteren Kompetenzbereich differenzierter werden“ sollen (Senatsverwaltung für Bildung, Jugend und Familie, 2010). Im folgenden Durchgang wurden die Testhefte vom Land Berlin verändert (Kuhl et al., 2011, S.18):

- Von den 22 Aufgaben im Bereich der Leitidee Zahlen und Operationen des Mathematiktests wurden zwei Aufgaben mit Lösungshäufigkeiten von 8% und 11% gegen zwei

Aufgaben aus 2008 mit Lösungshäufigkeiten von 58% und 56% ausgetauscht, wodurch sich die mittlere Lösungshäufigkeit dieses Testheftteils von 33,0% auf 37,7% vergrößerte.

- Die zweite Leitidee Muster und Strukturen büßte von Ihren 23 Aufgaben eine mit 14% Lösungshäufigkeit ein und wurde durch zwei Aufgaben mit 81% und 78% ergänzt, so dass die mittlere Lösungshäufigkeit für diesen Teil des Mathematiktests von 50,8% auf 54,7% stieg.

- Für den Lesetest stellte man fest, dass die mittlere Lösungshäufigkeit in der Pilotierung von 50,3% im Vorjahr auf 64,0% gestiegen war. Eine Anpassung darüber hinaus wurde nicht vorgenommen.

Dabei wurden bei der Berechnung der Kompetenzstufen die entsprechenden Itemparameter aus den Vortests genutzt, so dass sich lediglich die mittlere Lösungshäufigkeit veränderte, nicht aber die einer Leistung entsprechende Kompetenzstufenzuordnung (näheres dazu unter Abschnitt 3.3). Eine solche Anpassung wurde für das Land Berlin in keinem weiteren Jahr vorgenommen.

Eine Anpassung des Testhefts für jede Klasse war im ersten Durchgang der 2004 noch von der Universität Landau verantworteten Vergleichsarbeiten (noch in der Klassenstufe 4) gelebte Praxis. Etwa die Hälfte des Tests wurde zentral vorgegeben und die Lehrkräfte wählten sich dann einzeln Aufgaben hinzu. Dabei wurde das zur Wahl stehende Angebot an Aufgaben nach jeder ausgewählten Aufgabe angepasst. Auch, wenn zur Auswahl der Aufgaben keine Untersuchungen zu finden sind, kann doch angenommen werden, dass nicht nur die Schwierigkeit der Aufgaben, sondern ebenso inhaltliche Aspekte für die Auswahl berücksichtigt wurden. Durch dieses Verfahren wird, anknüpfend an die Idee des formativen Assessments, genau die Idee einer schul- bzw. klasseninterne Auseinandersetzung mit den Ergebnissen und eine Weiterarbeit auf diesen Ebenen unterstützt. Damit wird zudem der Weinert'schen Forderung der Sensibilität für individuelle Kontexte Rechnung getragen (Weinert, 2001a) und der Unterrichtsentwicklung wird mit dem Rückbezug der Daten auf den eigenen Kontext (Pant, 2011) mehr Raum gegeben.

Für VERA-8 wurden bis zum Jahr 2015 vom IQB für jede Domäne jedes Faches genau drei Testheftversionen für die Länder zur Verfügung gestellt, welche auf die Leistungserwartungen von Schülerinnen und Schülern im dreigliedrigen Schulsystem abgestimmt waren (siehe Tabelle 1.2). In Berlin wurde ab 2011 ein zweigliedriges Schulsystem eingeführt; neben dem Gymnasium gab es lediglich die Integrierte Sekundarschule (ISS)[7]. Die Bildungsverwal-

[7]Die Gemeinschaftsschule ist eine besondere Form der Integrierten Sekundarschule, die Primar- und Sekun-

tung legte jedes Jahr für jedes Fach fest, welches der zwei Testhefte der ISS und welches dem Gymnasium zugeordnet wurde. Im Allgemeinen wurde so verfahren, dass die zwei einfacheren Versionen A und B der ISS und dem Gymnasien zugeteilt wurden. Die schwierigste Version C wurde nur besonderen Gruppen zur Verfügung gestellt. So erhielten bilingual Deutsch/Englisch unterrichtete Klassen die Version C des Englischtests und Mathematik-Profilklassen den schwierigsten Mathematiktest. Mit der Reduktion auf zwei Testheftvarianten verblieb keinerlei Entscheidungsspielraum. Alle Gymnasien erhielten weiterhin das schwierige Testheft, eine anhand der Abbildung 4.5 nachvollziehbare Entscheidung.

1.6. Weiterentwicklung der Vergleichsarbeiten

Die Bestrebungen der KMK die Vergleichsarbeiten für die Schulpraxis attraktiver zu gestalten, führten zu einer Weiterentwicklung von VERA (KMK, 2018b), die kurz mit *Flexibilisierung* und *Modularisierung* beschrieben wird. Mit der Flexibilisierung verabschiedet man sich einerseits von bundesweit einheitlichen Testtagen. Die Tests finden seit dem in einem in der VERA-Steuergruppe zwischen den Ländern jährlich zu vereinbarendem Zeitkorridor von 4 bis 5 Wochen statt. Einige Länder stellen ihren Schulen diesen oder einen kleineren Korridor zur Verfügung, andere Länder, wie auch Berlin, halten weiterhin an einem für alle Schulen verbindlichen Testtermin fest. Wesentlicher aber ist, dass das Angebot an Testinstrumenten aufgefächert wird und dabei für jede Testdomäne mehrere Module bereitgestellt werden (Modularisierung). Diese unterscheiden sich teilweise auf fachinhaltlicher Ebene, teilweise in Bezug auf die Schwierigkeit. Ein Modul ist dabei ein Set aus Aufgaben, das in dieser Konstellation zusammen pilotiert/normiert worden ist. Die vom IQB in den jeweils zurückliegenden drei Jahren entwickelten Module können dabei in die Auswahl einbezogen werden. Inwieweit dieses, sich sukzessiv entwickelnde Angebot des IQB von den einzelnen Bildungsadministrationen der Länder tatsächlich dazu genutzt wird, den Schulen bzw. den Lehrkräften eine flexible Auswahl von Testinhalte zu ermöglichen, wird sich in den kommenden Jahren zeigen.

Sollen letztendlich auch Lehrkräfte eine Auswahl treffen dürfen, stellt sich die Frage, welche Rationalen könnten diese entwickeln, um eine gezielte Entscheidung für einen Test mit einer angemessenen Schwierigkeit zu treffen? Welche Informationen müssen die administrierenden Institute dafür zur Verfügung stellen? Solche Desiderate müssen mit wissenschaftlicher Forschung bearbeitet werden. Eine erste Arbeit zur möglichen Unterstützung bei der

darstufe I und ggf. eine Sekundarstufe II umfasst.

Testheftwahl durch Berliner Lehrkräfte liegt mit Weiß Aparicio (2021) vor. Die mit der Neuaufstellung von VERA verbundenen Veränderungen sollten dahingehend beobachtet werden, ob die intendierten Ziele auch erreicht werden.

2. Validität als integriertes, bewertendes Urteil

Die vorliegende Arbeit untersucht einige Aspekte der Validität am Gegenstand der Vergleichsarbeiten. Dabei bilden die in diesem Kapitel explizierten validitätstheoretischen Grundlagen den Rahmen, in den die Untersuchungen aus den drei zentralen Kapiteln der Arbeit abschließend eingeordnet werden.

2.1. Validität

Die Konzeptualisierung von Validität ist seit der Mitte des letzten Jahrhunderts fortwährend Diskussionen ausgesetzt (als kurzer Abriss siehe Frey, 2014; oder umfassender bei van den Ham, 2015). Der Rückblick von Messick (1989b) verweist auf die frühere Zergliederung der Validität in die drei Kategorien oder Arten der inhalts-, kriteriums- und konstruktbezogenen Validität. Wie andere Autor*innen auch beklagt er die damit verbundene Beliebigkeit der Verwendung

> „Furthermore, the continuing reference to three categories of validity evidence perpetuates, no less than did reference to three types of validity, the temptation to rely on only one (or, worse still, any one) category of evidence as sufficient for the validity of a particular test use.“ (ebenda, S.9)

und argumentiert letztendlich für ein integrierendes Konzept unter dem Dach der Konstruktvalidität. Inhalts- und Kriteriumsvalidität sind darin weiterhin Facetten der Validität. Messick macht aber deutlich, dass diese genauso wenig als unverbunden betrachtet werden können, wie Wertimplikationen und die Konsequenzen der Interpretation und Nutzung. Mit diesen letzten zwei Facetten erweitert er den Rahmen von Validität maßgeblich. Validität wird von Messick (1989a, S.13) wie folgt definiert[1]:

[1] hier zitiert aus der Übersetzung von Hartig et al. (2012, S.144)

> „Validität ist ein integriertes, bewertendes Urteil über das Ausmaß, in dem die Angemessenheit und Güte von Interpretationen und Maßnahmen auf der Basis von Testwerten [...] durch empirische Belege und theoretische Argumente gestützt sind."

Diese Konzeptualisierung kann mit der Aufnahme in die Empfehlungen von American Educational Research Association (AERA) et al. (2014) als weitgehender, wenn auch nicht unwidersprochener Konsens angesehen werden.

Andere Autoren folgen der Idee eines ganzheitlichen Validitätskonzeptes nicht bzw. nicht in dieser umfassenden Form. Gegenentwürfe reduzieren die Validität (wieder) auf eine Eigenschaft des Tests (Borsboom et al., 2004; Lissitz & Samuelsen, 2007). Bei Lissitz und Samuelsen (2007) konstituiert sich die zu messende Eigenschaft durch die Operationalisierung, also durch die verwendeten Items, Verknüpfungen zu Definitionen anderer Tests oder anderer Domänen sind für sie nicht notwendig. Die Definition von Validität fokussiert damit im Kern die Entwicklung des Tests. Weiter erkennen einige (wie auch Lissitz & Samuelsen, 2007) Testwertinterpretation und -nutzung zwar als relevante Untersuchungsgegenstände an, aber gerade nicht als Teil der Feststellung von Validität. Dass der Aspekt der konsequenzbezogenen Validität ein strittiger Diskussionspunkt ist, hat auch Kane (2013, S.60) schon hervorgehoben. Der Abschnitt 2.2 stellt die Bedeutung dieses Aspekts noch einmal heraus.

Der argumentbasierte Ansatz von Kane (2013) überführt diese theoretischen Überlegungen in einen praktisch nutzbaren Rahmen. Die Feststellung von Validität erwartet dabei ein *Validierungsargument* (ebenda, S.14ff), welches die Interpretation und Verwendung der Ergebnisse in kohärenter Weise durch verschiedene theoretische Annahmen und empirische Belege stützt. Als theoretischen Rahmen für die Argumentation(en) bezieht sich Kane auf frühe Arbeiten von Toulmin (2003)[2] und sein darin entwickeltes Modell von Schlussfolgerungen.

In der Tabelle 2.1 sind die von Kane (2013) am typischen Verlauf von Assessments orientierten, sich sukzessive aufeinander aufbauende *Schlussfolgerungen* den entsprechenden *Validitätsaspekten* zugeordnet, die mit Bezug auf Vorarbeiten von Messick von Schaper (2014, S.26) zusammengefasst wurden. In der Auseinandersetzung mit den Schlussfolgerungen in der Tabelle 2.1 werden aber auch zwei jeweils zusammenhängende Positionen deutlich. Wird ein Test im ersten Schritt derart konstruiert (1)[3], dass durch die Operationalisierung eine angemessene Modellierung der Zieldomäne gewährleistet ist, dass also inhaltliche Validität

[2] Diese Version des Buches von Toulmin ist die überarbeitete Auflage seiner ersten Veröffentlichung von 1958.
[3] Diese und folgende in Klammern gesetzten Nummern beziehen sich auf die Nummerierung der Prozessschritte in der Tabelle 2.1.

hergestellt wird, dann ist dies das grundlegende Argument dafür, dass eine Extrapolation (5) des erwarteten Testergebnisses auf die Feststellung entsprechend vorliegender Kompetenz in der Zieldomäne funktioniert. In gleicher Weise sorgen standardisierte Bedingungen (2) bei der Testdurchführung dafür, dass das Testergebnis auch über die Messung in äquivalenten Situationen generalisiert (4) werden kann.

Pant et al. (2010, S.178) verweisen auf die Notwendigkeit „zwischen dem Validierungsargument für das System des Large-Scale-Assessments als Ganzes und dessen Subsystemen“ zu differenzieren. Um beispielsweise der Zuordnung der Leistungsfähigkeit von Schüler*innen zu Kompetenzstufen auf der Basis ihrer Testergebnisse Validität zu bescheinigen, zählen Pant et al. (2017) fünf Argumente auf, deren Gültigkeit dafür zu belegen wären. In der Folge untersuchen sie mit der Festlegung begründeter Cut-Scores eines dieser Argumente, welches sie als das schwächste Glied der Argumentation herausstellen. Andere Autor*innen entwickeln hingegen die gesamte Argumentationskette für etablierte Tests (Chapelle et al., 2010, für den TOEFL; van den Ham, 2015, für einen Mathematiktest des NEPS) und zeigen dabei, wie verschiedene Prozessschritte der Entwicklung, Anwendung und Nutzung von Tests jeweils eigener Argumente bedürfen, um die letztendlichen Interpretationen und Nutzungen als valide herauszustellen. Deutlich wird dabei die Relevanz jedes einzelnen Schrittes für die gesamte Argumentationskette. Der Entwurf von Kane (2013) wird dabei, wie von ihm selbst inspiriert, adaptiv erweitert. So wird die Skalierung bei van den Ham (2015) aus der Bewertung herausgelöst und separat behandelt.

2.2. Die Bedeutung konsequenzbezogener Validität

Kane (2013) und letztendlich auch American Educational Research Association (AERA) et al. (2014) verweisen darauf, dass der Nachweis der Validität insbesondere auch die Konsequenzen einzubeziehen hat. So wird in den Standards for Educational an Psychological Testing (American Educational Research Association (AERA) et al., 2014, S.24, Standard 1.5) erklärt:

> „When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.“

Die Standards führen sogar weiter aus, dass die Gründe für die Erwartung auch des indirekten Nutzens explizit dargelegt und mit theoretischen Argumenten wie empirischen Belegen

Tabelle 2.1.: Gegenüberstellung der Schlussfolgerungen bei Kane und der Validitätsaspekte bei Messick in einem typischen Assessmentverlauf

	Assessmentschritt	Schlussfolgerung[a]	Validitätsaspekt[b]
1	Domänenbeschreibung/ Testkonstruktion		Inhaltliche Validität: Curriculare und theoretische Absicherung des modellierten Bereichs
2	Testdurchführung	die Schlussfolgerung von der Bearbeitung der einzelnen Aufgaben zu einem (beobachteten) Testergebnis	Kognitive Validität: Passung der kognitiven Prozesse bei der Kompetenzerfassung zum postulierten theoretischen Kompetenzmodell
3	Bewertung/ Skalierung	die Schlussfolgerung vom beobachteten Testergebnis auf das in äquivalenten Testsituationen erwartbare Testergebnis	Strukturelle Validität: Passung von theoretischem Kompetenzmodell und gewähltem psychometrischem Messmodell
4	Generalisierung	die Schlussfolgerung von diesem erwarteten Testergebnis auf die Kompetenz in der dem Test zugrunde liegenden Zieldomäne	Angemessenheit einer über die Aufgaben- und Personengruppe hinausgehenden Interpretation
5	Extrapolation	die Schlussfolgerung von auf die zugeschriebenen Kompetenz rekurrierenden Entscheidungen	Angemessenheit mit Blick auf konvergente, diskriminante und prädiktive Zusammenhänge mit anderen Konstrukten
6	Interpretation/ Entscheidung		Konsequentielle Validität: Angemessenheit mit Blick auf die Interpretation der Ergebnisse und die Konsequenzen der Schlussfolgerungen

[a] entnommen aus Kane (2013).
[b] entnommen aus Schaper (2014).

begründet werden sollen (ebenda, S.24, Standard 1.6), wobei entgegenstehende Befunde angemessen zu berücksichtigen sind. Zur Frage, wem diese Begründungspflicht aufzuerlegen ist führt Kane (2013, S.62) aus, dass einerseits diejenigen, die über den Einsatz eines Tests entscheiden die Verantwortung dafür tragen, dass deklarierte Ziele mit dem Einsatz des Tests erreicht werden können und dass andererseits die Entwickler*innen von Tests einen Teil der Verantwortung dafür übernehmen, indem sie belegen, dass der Test dazu geeignet ist. Hinzu kommt, dass

> „... test users generally are in the best position to identify unintended consequences, but test publishers also have responsibility for the consequences of uses they explicitly or implicitly advocate“ (ebenda, S.58).

Gefordert ist demnach sowohl eine ex ante begründete Annahme der Zielerreichung sowie ein ex post zu erbringender Nachweis.

Validität ist eine Eigenschaft, die nicht dem Testinstrument allein sondern dem gesamten Prozess der Nutzung in gradueller Ausprägung zugeschrieben wird. Dieser Nachweis erfordert eine fortwährende Überprüfung jeder Art von Testwertinterpretation und -nutzung in der Form, dass die Gültigkeit des Arguments, auf das sich eine konkrete Interpretation und Nutzung bezieht, durch theoretische Argumente und empirische Belege gestützt wird. Für zwei differente Nutzungen sind die Nachweise entsprechend einzeln zu erbringen. Das *Validitätsargument* nach Kane meint eine Argumentationskette, die aus einzelnen Schlussfolgerungen besteht (vergleiche Tabelle 2.1). Für jede dieser Schlussfolgerungen gilt es ein Argument zu formulieren, welches die Schlussfolgerung begründet und dieses Argument wird dann wiederum auf theoretisch und empirisch zu prüfende Annahmen gestützt. Der folgende Abschnitt skizziert, wie eine solche Argumentation entsprechend dem Modell von Toulmin (2003) erfolgt.

2.3. Das Modell der Argumentation von Toulmin

Ausgangspunkt einer Schlussfolgerung ist bei Toulmin (2003) ein *Datum*, also ein Wert bzw. eine Aussage, auf die sich eine Schlussfolgerung (*Claim*) gründet. Diese Schlussfolgerung stützt sich auf ein Argument (*Warrant*), dessen zugrunde liegende Annahmen durch empirische Untersuchungen belegt und durch theoretische Argumente begründet werden sollen. Die Schlussfolgerung kann zudem durch sogenannte *Qualifier* beeinflusst werden. Modale Qualifier beschreiben, in welchem Grad die Schlussfolgerung zur Behauptung führt, andere

beschreiben Ausnahmen (Exceptions), unter denen die Schlussfolgerung falsch ist[4].

Um aus der Argumentationskette ein Beispiel herauszugreifen, könnte man die Schlussfolgerung betrachten, mit der von einem Antwortvektor, der für jede Aufgabe eines Mathematiktests die jeweilige Kodierungen für richtig bzw. falsch gelöste Items enthält, auf ein beobachtetes Ergebnis als Punktzahl geschlossen wird. Dabei ist es gleichgültig, ob sich das Ergebnis als relative Häufigkeit der richtig gelösten Aufgaben ergibt oder als Logit nach einer Rasch-Skalierung. Das Verfahren beschreibt die Regel bzw. beantwortet die Frage, wie man von den Daten auf das Ergebnis schlussfolgert. Dazu muss gezeigt werden, dass die mit der Schlussfolgerungen verbundenen Annahmen zutreffend sind. Diese Annahmen sind für die Bildung einer einfachen relativen Häufigkeit einerseits und die Anwendung einer Rasch-Skalierung andererseits teilweise ähnlich, teilweise aber auch unterschiedlich. Die Rasch-Skalierung nutzt dabei Itemparameter, die wiederum Ergebnis (also Schlussfolgerung) anderer Prozesse sind, so dass sich quasi ein „never-ending process“ (Kane, 2013, S.15) von Schlussfolgerungen und Annahmen entspinnen würde. Einige grundlegende Annahmen werden deshalb als a priori ausreichend plausibel betrachtet. Die Feststellung von Validität ist demnach ein Prozess der versucht, insbesondere für die fragwürdig(st)en Annahmen plausible Belege zu finden und so eine kohärente Argumentation entwickelt.

2.4. Das Validitätsargument für die Vergleichsarbeiten

Für die in dieser Arbeit im Fokus stehenden Vergleichsarbeiten stellt sich die Frage, in wieweit die Interpretationen und Nutzungen insbesondere der kompetenz(stufen)bezogenen Rückmeldungen als Ergebnis der Testkonstruktion und der Testanwendung als valide gelten können. Mit der Unterstützung von Unterrichtsentwicklung geht die Zielstellung der Vergleichsarbeiten deutlich über die reine Feststellung eines Lernstandes hinaus. Als Rahmen der vorliegenden Arbeit wurde deshalb der argumentative Ansatz von Kane (2013) gewählt und damit die Aspekte der Testwertinterpretation und der Konsequenzen eingeschlossen.

In diesem Abschnitt wird das Validitätsargument für die Vergleichsarbeiten entsprechend der bei Kane zu Grunde liegenden Phasen der Durchführung skizziert. Diese Skizze soll keine vollständige Argumentationskette entwickeln, sondern lediglich dazu dienen, die Beiträge der vorliegenden Arbeit einzuordnen. Von besonderer Relevanz bei den Vergleichsarbeiten ist aber, dass die Schlussfolgerungen teilweise auf Bedingungen rekurrieren, die sich aus vorgängigen Prozessen ergeben, die nicht Teil der eigentlichen VERA-Durchführung sind. Einer-

[4] Bei Kane (2013) und auch anderen Autor*innen werden die verschiedenen Qualifier zusammengefasst.

seits wird durch die Normierung der Bildungsstandards die jeweils domänenbezogene Skala der Bildungsstandards definiert, an welche sich die Ergebnisse der Vergleichsarbeiten direkt anbinden. Mit dem Standard-Setting wird andererseits diese Skala zur Interpretation in inhaltlich definierte Stufen eingeteilt. Erst über die aus der Pilotierung der VERA-Aufgaben resultierenden Item-Parameter können die VERA-Ergebnisse auf diese Skala und ihre Kompetenzstufen bezogen werden. Diese Verknüpfungen sind von herausgehobener Bedeutung und deshalb in der vorliegenden Skizze berücksichtigt.

Bewertung

Nach der Bearbeitung der einzelnen Aufgaben eines Testhefts, werden die Antworten im Allgemeinen durch die Lehrkraft bezüglich ihrer Korrektheit bewertet. Das Ergebnis dieser Bewertung der Antworten der Schüler*innen wird als Antwortvektor bezeichnet. Für diese erste Schlussfolgerung ist zu zeigen, dass die Bewertung zu einem Antwortvektor führt, der die Fähigkeit der korrekten Beantwortung im Sinne der Testentwickler*innen tatsächlich abbildet.

Skalierung

Mit der sich anschließenden Rasch-Skalierung wird vom diesem Antwortvektor auf ein beobachtetes Testergebnis geschlossen, welches als Fähigkeitswert auf einer latenten Skala (Skala der Bildungsstandards) beschrieben wird. Um die Angemessenheit der Skalierung festzustellen, sind verschiedene Bedingungen und Voraussetzungen zu formulieren und zu überprüfen, die sich aus den weiter unten beschriebenen Anforderungen für den Einsatz der Rasch-Skalierung ergeben. Grundlegend ist dabei, dass die Skalierung Schwierigkeitsparameter für sämtliche Aufgaben einbezieht, die das Ergebnis einer Pilotierung/Normierung sind, die über Ankeraufgaben eine Verknüpfung (Linking) mit der Metrik der Bildungsstandards herzustellen versucht. Da das Testergebnis der Vergleichsarbeiten auf dieser Metrik interpretiert wird, muss für die Schlussfolgerung notwendig die Korrektheit der Schwierigkeitsparameter als Ergebnis des Linkings gelten.

Generalisierung

Die folgenden zwei Schlussfolgerungen nehmen Verallgemeinerungen vor. Die Erste soll deutlich machen, dass das beobachtete Testergebnis über die konkrete Testung hinaus ein für die Fähigkeit erwartbares Ergebnis darstellt. Denn mit der Interpretation wird unterstellt, dass die Schülerin oder der Schüler das gleiche Ergebnis unabhängig von der eingesetzten Test-

version, also auch mit anderen Aufgaben, unabhängig von den konkreten Testbedingungen gezeigt hätte. So zeigt van den Ham (2015) hier, dass die Testung „sorgfältig anhand von angemessenen, standardisierten Prozeduren durchgeführt“ wurde und dass „die individuellen Fähigkeitsmessungen reliabel“ sind.

Extrapolation

Die zweite das Testergebnis verallgemeinernde Schlussfolgerung bezieht sich auf die inhaltliche Bedeutung der Messung, denn das Testergebnis wird im Allgemeinen als Ausprägung der Kompetenz in der Zieldomäne interpretiert und nicht nur als das Ergebnis eines Tests. Verfahren, mit denen konvergente oder diskriminante Validität gezeigt wird, können entsprechende Argumente für eine solche Extrapolation stützen. Natürlich ist eine angemessene Testkonstruktion maßgeblich dafür, dass dies gelingt.

Im Fall der Nutzung von Kompetenzstufen in der Ergebnisdarstellung schließt sich eine weitere Schlussfolgerung an. Diese bezieht sich auf eine im Allgemeinen vorgängig festgelegte Beschreibung von Kompetenzniveaus. Hierbei wird die zu belegende Annahme unterstellt, dass die Niveaus beschreibenden inhaltlichen Fähigkeitsaspekte durch entsprechende Testaufgaben repräsentiert sind, zu deren Lösung genau diese Fähigkeiten notwendig sind. Überdies muss dann die korrekte Lösung einer entsprechenden Aufgabe mit hinreichender Wahrscheinlichkeit zur Zuordnung von Testleistungen in der entsprechenden Stufe führen.

Auch hier wird die Argumentation mit Daten und Schlussfolgerungen verknüpft, die nicht Teil des Prozesses der Vergleichsarbeiten selbst sind. Die Operationalisierung ist wesentlich das Ergebnis einer Aufgabenauswahl, die sich auf die Ergebnisse der Pilotierung/Normierung stützt. Die Beschreibung der Kompetenzstufen durch Expert*innen liegt wiederum dem Prozess des Standard-Settings zugrunde, der festlegt, wie die kontinuierliche Skala in Niveaustufen unterteilt wird (Pant et al., 2017). Demnach ist die Validität der Pilotierung/Normierung sowie des Standard-Settings äquivalent nachzuweisen, um das Valditätsargument für die Vergleichsarbeiten zu belegen.

Interpretation und Entscheidung

Für die Interpretation der Ergebnisse ist von großer Bedeutung, ob sich diese auf eine einzelne Person bezieht oder auf das Ergebnis einer größeren Gruppe von Schüler*innen, vielleicht sogar auf eine Aussage über die Kompetenzstufenverteilung im ganzen Land. Welche Interpretation fokussiert wird, hat aber auch Rückwirkungen auf Schlussfolgerungen vorhergehender

Schritte. So ist eine Kompetenzstufenzuordnung für eine einzelne Person von herausragender Unsicherheit, für die Verteilung im ganzen Land wird allgemein eine große Stabilität angenommen. Der Nachweis von Validität ist explizit für die Interpretation einer Art von Ergebnissen zu führen. Wird nicht die Zuordnung der Kompetenzstufe für einzelne Schüler*innen betrachtet sondern für das ganze Land oder steht die prozentuale Lösungshäufigkeit der Aufgaben im Zentrum des Interesses, sind die einzelnen Schritte des Validitätsnachweises erneut zu untersuchen. In einigen Fällen können dabei einzelne Schlussfolgerungen mit identischen Argumenten begründet werden.

Die vorliegende Arbeit fokussiert die Betrachtung der Zuordnung von Kompetenzstufen für Schülerinnen und Schüler. Das teilweise die Verteilung von Kompetenzstufen für die Länder Berlin und Brandenburg und hier teilweise auch nur für die Gymnasien untersucht werden, unterstützt lediglich eine Argumentation für spätere Schlussfolgerungen für die Messung auf Ebene der Einzelschüler*innen. Überdies vertritt der Autor die Ansicht, dass auch die Verteilung von Kompetenzstufen für eine Klasse lediglich eine Kumulation von Kompetenzzuordnungen einzelner Schülerinnen und Schüler darstellt. Schlussfolgerungen einer Lehrkraft, die nicht danach fragen, wer die einzelnen Schüler*innen sind die beispielsweise nicht die Mindeststandards oder schon die höchste Kompetenzstufe erreichen, sollten in der Praxis eine Ausnahme darstellen.

2.5. Einordnung der Beiträge der vorliegenden Arbeit

Die vorliegende Arbeit entstand aus der fortwährenden Beobachtung und Gestaltung der Vergleichsarbeiten, insbesondere mit Blick auf den Vollzug in der Praxis. Die linke Seite der Abbildung 2.1 ist eine Skizze des Validitätsarguments für die Vergleichsarbeiten nach Kane. Berücksichtigt ist, wie auch in den vorgängigen Beschreibungen der Schlussfolgerungen, die Verknüpfung mit der VERA-Pilotierung (Mitte). Diese liefert für die Skalierung im VERA-Durchgang die Itemparameter. Die Validität der VERA-Pilotierung ist damit eine grundlegende Stütze für die Argumentationskette des VERA-Durchgangs. Die VERA-Pilotierung nutzt zum Zwecke der Verlinkung der neuen VERA-Aufgaben mit der Skala der Bildungsstandards Aufgaben aus der Normierung (Rechts), womit auch diese implizit Bedeutung für die Argumentation hat. Diese beiden Verknüpfungen sind von besonderer Tragweite, weil erst damit der Bezug der Ergebnisse des VERA-Durchgangs auf die Beschreibungen der Zieldomäne und der damit verbundenen Kompetenzstufenmodelle möglich wird.

Die drei Kapitel 4, 5 und 6, welche in der Abbildung 2.1 verschiedenen Knoten der Ka-

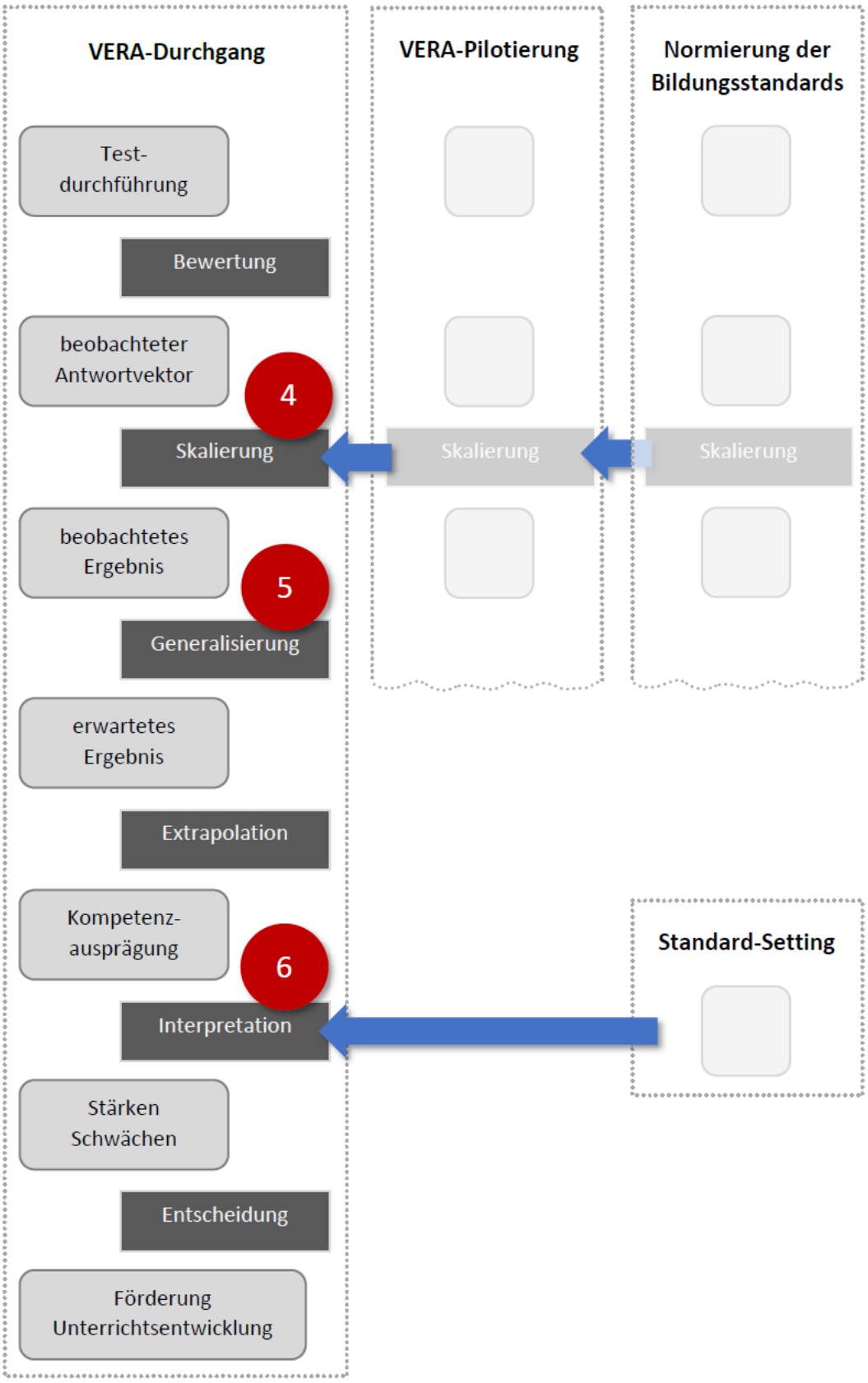


Abbildung 2.1.: Skizze des Aufbaus eines Validitätsarguments für die Vergleichsarbeiten nach Kane und Auszeichnung der Schlussfolgerungen, für die in den Kapiteln 4 bis 6 Belege gesucht werden

ne'schen Argumentationskette zugeordnet sind, sind als einzelne Beiträge zur Validitätsdiskussion zu betrachten.

Im Kapitel 4 *Überprüfung von Gewissheiten beim Einsatz der Rasch-Skalierung* werden die Bedingungen der Rasch-Skalierung als wesentlicher Baustein struktureller Validität betrachtet. Hierbei wird insbesondere auf solche Bedingungen eingegangen, welche die konkrete Fähigkeitsschätzung auf der Basis des Antwortvektors und damit die Interpretation des Fähigkeitswertes betreffen. Überdies wird die Bedeutung für die Zusammenstellung von Testheften untersucht. Die einzelnen Betrachtungen sollen Schlussfolgerungen aus dem Bereich der *Skalierung* stützen.

Für den Schritt der Generalisierung im Validitätsargument soll im Kapitel 5 *Stabilität der Ergebnisse von Vergleichsarbeiten* gezeigt werden, dass die erwarteten Ergebnisse über die Testsituation und dabei konkret über das eingesetzte Testheft hinaus generalisiert werden können. Um dies zu zeigen, werden die Ergebnisse vergangener VERA-Messungen mit unterschiedlichen Testinstrumenten und Messungen anderer bildungsstandardbezogener Tests gegenübergestellt und mit entsprechenden Erwartungen abgeglichen. Gelingt diese Argumentation, belegt sie die korrekte und belastbare Schlussfolgerung vom beobachteten auf das erwartete Testergebnis.

Die oben ausgeführte Notwendigkeit der Überprüfung der Zielerreichung begründet die Bedeutung der Untersuchung der Ausschöpfung von Rückmeldeabrufen im Kapitel 6 *Vor der Rezeption*. Diese kann als Qualifier (Exception) für die Schritte *Interpretation* und *Entscheidung* verstanden werden. Sämtliche Aussagen zur Nutzung der Ergebnisrückmeldungen von Vergleichsarbeiten, wie sie in vielen Studien der zurückliegenden Jahre getätigt wurden, sind nur unter der Bedingung gültig, dass die Abrufe in angemessenem Maße erfolgen. Bleiben Ergebnisse ungenutzt, sind mindestens die ohne jegliche Folgen eingesetzten finanziellen und zeitlichen Ressourcen als nicht intendierte negative Folgen aufzuführen. Für die Bewertung konsequentieller Validität ist dies dann von Bedeutung, wenn solche Nicht-Nutzung keine Ausnahme darstellt, sondern sich als systematisch erweist. Dieses Kapitel nähert sich diesem bisher kaum betrachteten Aspekt der Abrufe von Rückmeldungen explorativ und entwickelt dabei neue Kriterien für eine Bewertung des Gegenstandes.

3. Raschskalierung, Standardsetting und Linking

Insbesondere die Untersuchung der ersten zwei Fragestellungen (Kapitel 4 und 5) rekurrieren auf einige psychometrische Grundlagen, die in diesem Kapitel erläutert werden. Die verschiedenen bei der Entwicklung von Tests relevanten Verfahrensschritte werden zuerst allgemein und dann für die konkrete Umsetzung der Vergleichsarbeiten beschrieben.

Ausgangsbasis für die Konstruktion eines Tests ist die Festlegung a) des zu messenden Konstrukts, b) der Grundgesamtheit für die diese Messung ausgelegt sein soll sowie c) des Ziels der Ergebnisnutzung. Im Fall der mit den Vergleichsarbeiten verbundenen Zielstellung der Überprüfung der KMK-Bildungsstandards, liegen mit diesen Standards zugleich konkrete Beschreibungen der Konstrukte vor, die zudem auf eine konkrete Zielpopulation bezogen sind. Mit der Festlegung der Durchführung als schriftliche Tests, beschränken sich die im Rahmen der Vergleichsarbeiten vorgenommenen Messungen auf die im vorgehenden Kapitel beschriebenen Kompetenzen. Die Entwicklung jedes einzelnen Instruments für jede Kompetenz erfordert dabei bestimmte Prozessschritte (ein Vergleich dieser zwischen Ländervergleich[1] und VERA findet sich bei Köller, 2008).

Zuerst werden auf der Basis von Kompetenz- und Kompetenzstufenbeschreibungen Aufgaben entwickelt. Dazu werden geeignete Personen rekrutiert und zwischen diesen die Anforderungen für die Aufgabenentwicklung abgestimmt (Itzlinger-Bruneforth et al., 2016, ab S.31). In der Prä-Pilotierung werden die Aufgaben einer kleinen Stichprobe von Schüler*innen vorgelegt, um so die grundsätzliche Eignung der entwickelten Aufgaben zu überprüfen. Der Fokus liegt dabei auf der Konstruktebene, also der Frage: Funktionieren die Aufgaben so, wie die Entwickler*innen sich das fachdidaktisch überlegt hatten? Die erfolgreich prä-pilotierten Aufgaben werden dann in einer Pilotierung bezüglich ihrer psychometrischen Eignung überprüft. Dazu muss die Stichprobe alle relevanten Anteile der Grundgesamtheit erreichen. Beispielsweise überprüft die Pilotierung, ob alle Aufgaben in allen Teilgruppen die zu messende

[1] später Bildungstrend

Kompetenz in der gleichen Weise abbilden und ob die Distraktoren von Auswahl-Aufgaben sinnvoll gewählt sind, um die Kompetenz trennscharf zu messen. Die fachinhaltlich wie psychometrisch besten Aufgaben verbleiben im Aufgabenpool, der auf eine für die Messung als notwendig bestimmte Zahl von Aufgaben reduziert wird. Unter 3.2 finden sich einige Überlegungen zu dieser Auswahl.

An einer repräsentativen Stichprobe erfolgt die Normierung der Testhefte, die aus den ausgewählten Aufgaben zusammengestellt worden sind. Das Ergebnis der Normierung ist eine Metrik für die zu messende Kompetenz. Weil die Basis aller hier beschriebenen Entwicklungen die Bildungsstandards sind, wird diese Metrik deshalb auch hier als BiSta-Metrik bezeichnet. Die Metriken sind für die einzelnen Kompetenzen natürlich eigenständig festgelegt, die Messwerte für das Hörverstehen in Englisch und die Kompetenzmessung in Mathematik nicht direkt aufeinander beziehbar. Um darüber hinaus auch eine kriterielle Bewertung der gemessenen Fähigkeit zu ermöglichen, werden auf dieser Metrik im Rahmen eines Standard-Setting-Prozesses durch Cut-Scores voneinander abzugrenzende Fähigkeitsbereiche als Kompetenzstufen definiert. So kann beispielsweise festgestellt werden, dass eine Person für das Leseverstehen in Englisch einen Regelstandard erreicht, nicht aber für Mathematik. Dieser Bezug wird aber eben nur mittelbar über die Festlegungen im Standard-Setting auf der zuvor festgelegten Metrik definiert.

Nach allgemeinen Bemerkungen zum Messen (3.1) werden die Prozessschritte der Operationalisierung (3.2), der Skalierung (3.3) sowie des Standard-Settings (3.4) in den folgenden Unterabschnitten in der für das Verständnis alles Folgenden notwendigen Tiefe beschrieben und ggf. Verweise zu detailreicheren Ausführungen gegeben. Die zwei Abschnitte 3.5 *Testdesign* und 3.6 *Linking* widmen sich den für Testwiederholungen wichtigen Verfahrensbestandteilen, deren Umsetzung im Rahmen der Vergleichsarbeiten (3.7) schließt das Kapitel ab.

3.1. Messen

Das Ergebnis eines quantitativen Tests wird üblich als numerisches Relativ auf einer Skala abgebildet. Dabei stehen im Allgemeinen höhere Werte für eine größere und kleinere Werte für eine niedrigere Leistungsfähigkeit. Damit wird die gegenstandsbezogene Leistungsfähigkeit als ein eindimensionales Merkmal definiert. Weiter wird unterstellt, dass Aufgaben mit dem Merkmal der Schwierigkeit unterschiedlich vorhandene Ausprägungen der Leistungsfähigkeit in der Art repräsentieren, dass die Fähigkeit einer Person, eine Aufgabe korrekt zu lösen, in einer Relation zur Schwierigkeit der Aufgabe steht. Damit verbunden ist die Vorstellung, dass

sich Schwierigkeit wie Fähigkeit auf einer Skala derart gemeinsam darstellen lassen, dass eine Testperson mit der korrekten Lösung einer Aufgabe im Allgemeinen eine Leistungsfähigkeit unter Beweis gestellt hat, die vom Maß her über der Schwierigkeit der Aufgabe liegt. Gelegentlich wird hier das Bild eines Hochspringers bemüht, dessen Fähigkeit sich an der Höhe der Latte, als Äquivalent zur Aufgabenschwierigkeit, beweist.

Bei einer sehr einfachen und oft verwendeten Form der Skalenkonstruktion, wird für jede richtig gelöste Aufgabe ein Punkt vergeben. Eine solche Skala erlaubt zunächst lediglich die Beschreibung relativer Positionen (Ordinalskala) und ein grundlegendes Problem dieser Art der Skala wird deutlich, wenn man sich zwei Tests vorstellt, von denen der eine nur einfache, der andere nur schwierige Aufgaben enthält. Für die identische Leistung ergeben sich verschiedene Maße auf den zwei Skalen, bzw. muss allgemein davon ausgegangen werden, dass jedes Set von Aufgaben eine andere Skala definiert. Die Idee, dass die Differenz zwischen der Fähigkeit einer Person und der Schwierigkeit einer von dieser Person bearbeiteten Aufgabe bewertet wird, führt zur Konstruktion eines Tests, der eine Bewertung der Fähigkeit unabhängig von der Auswahl der Testaufgaben ermöglicht.

Oft will man mit dem Ergebnis einer Fähigkeitsmessung eine Aussage darüber treffen, in welchem Ausmaß der durch den Test inhaltlich repräsentierte Fachinhalt beherrscht wird. So kann man schon bei der Testkonstruktion fachdidaktisch, kriterial begründet festlegen, ab welcher Aufgabenschwierigkeit der Test als *bestanden* gilt oder mehrere Grenzen festgelegt, welche die Zuordnung von unterschiedlichen Stufen als Grad der Ausprägung der gemessenen Fähigkeit begründen. Ist eine solche Stufenzuordnung fachinhaltlich begründet, ist sie ein akzeptierter Ausgangspunkt von Konsequenzen wie die Zuweisung von Förderung oder Schullaufbahnentscheidungen. Damit solche Schlussfolgerungen aus Messungen eine valide Grundlage haben, muss die Abbildung der Fähigkeiten auf die Messskala eindeutig sein und die Anbindung der inhaltlichen Beschreibung wohlbegründet erfolgen.

Eine letzte Überlegung führt zur Skalierung nach der von Georg Rasch (1960) begründeten probabilistischen Testtheorie (auch Item-Response-Theory, IRT). Bisher wurde davon ausgegangen, dass eine Person sämtliche Aufgaben, deren Schwierigkeit die eigene Fähigkeit überschreiten, bei denen also die Differenz von Personenfähigkeit und Aufgabenschwierigkeit negativ ist, sicher nicht gelöst werden und umgekehrt, dass alle Aufgaben, deren Schwierigkeit niedriger ist als die Fähigkeit, sicher gelöst werden. Statt dessen wird bei der Rasch-Skalierung davon ausgegangen, dass die Wahrscheinlichkeit für die korrekte Lösung einer Aufgabe dann genau 50% beträgt, wenn die Personenfähigkeit gleich der Aufgabenschwierigkeit ist. Für

schwierigere Aufgaben ist die Wahrscheinlichkeit für eine korrekte Lösung kleiner und für leichtere Aufgaben größer. Verfahren der IRT sind inzwischen im Rahmen von large scale assessments etabliert (OECD, 2003; OECD, 2017; Becker et al., 2019) und auch bei den Vergleichsarbeiten das Mittel der Wahl (zuletzt Aneis et al., 2020). Gegenüber den Verfahren der Klassischen Testtheorie (KTT) besitzen diese Verfahren einige Vorteile (Rost, 2004). Mit dem Standard-Setting (Pant et al., 2010) liegt zudem ein objektiviertes Verfahren für eine inhaltlich begründete Stufenbildung vor.

3.2. Operationalisierung

Leistungstests geben im Allgemeinen vor, den Grad der Ausprägung einer bestimmten Fähigkeit bzw. einer Kompetenz auszuweisen. Die Aufgaben eines Tests weisen dazu unterschiedliche Schwierigkeiten auf. Löst eine Person eine Aufgabe richtig, stellt sie damit eine Fähigkeit unter Beweis, die in einem Verhältnis zur Schwierigkeit der Aufgabe steht. Mit einer einzigen Aufgabe könnte man die gemessene Fähigkeit lediglich als hoch oder niedrig klassifizieren, wobei die konkrete Schwierigkeit der Aufgabe den Orientierungspunkt bestimmt. Üblich soll ein Test aber darüber hinaus messbar machen, in welchem Grad eine bestimmte Fähigkeit ausgeprägt ist. Dazu sind dann offensichtlich Aufgaben unterschiedlicher Schwierigkeit notwendig. Ein Test wird also durch mehrere angemessen ausgewählte Aufgaben operationalisiert. Durch die Auswahl von Aufgaben verschiedener Schwierigkeit wird eine Differenzierung der Ergebnisse möglich.

Das weiter oben erwähnte Beispiel zweier Tests, deren Aufgaben so einfach bzw. so schwer sind, dass eine Person alle bzw. keine korrekt zu lösen in der Lage ist, verdeutlicht, dass für eine differenzierte Messung der Fähigkeit eine bezüglich der Aufgabenschwierigkeiten angemessen differenzierte Auswahl notwendig ist. Dies bedeutet einerseits, dass eine entsprechende Zahl an unterschiedlich schwierigen Aufgaben benötigt wird sowie dass diese Schwierigkeiten dort liegen sollen, wo auch die Fähigkeiten der Personen vermutet werden. Diese Rücksichtnahme findet allgemein viel Beachtung. Im Rahmen der Vergleichsarbeiten werden deshalb Tests zur Verfügung gestellt, bei denen die Verteilung der Aufgabenschwierigkeiten die vermutete Verteilung der Personenfähigkeiten gut abbilden (siehe auch Abschnitt 1.5).

Dem gegenüber weniger thematisiert wird die angemessene Repräsentation der Teilfähigkeiten. Zuerst muss die zu messende Fähigkeit eindeutig beschrieben werden, damit auf dieser Basis eine angemessene Auswahl von Aufgaben erfolgen kann. Die so messbar gemachte Fähigkeit ist in dem Sinne nicht trivial, dass sie im Allgemeinen aus Teilfähigkeiten bzw.

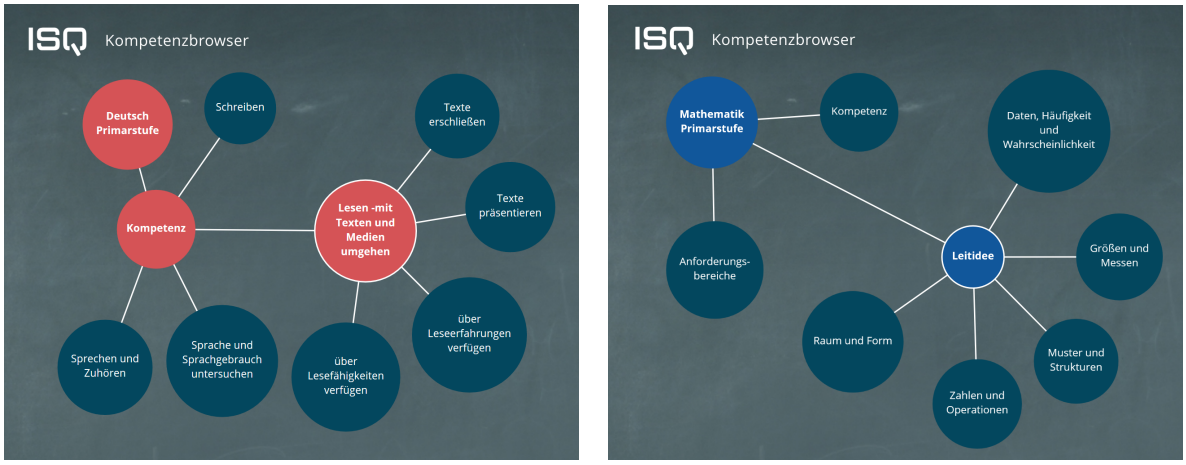


Abbildung 3.1.: Kompetenzmodell Primarstufe Deutsch (links) und Mathematik (rechts)

bestimmten Facetten besteht oder dass sie sich in jeweils bestimmten Teilbereichen ausprägt. So weisen die Bildungsstandards der Primarstufe für die Domäne *Deutsch Lesen* unterschiedliche Teilfähigkeiten auf, die ihrerseits weiter unterteilt sein können. Auch die Kompetenz *Mathematik* in der Primarstufe wird im IQB-Bildungstrend gemessen. Hier werden in den Bildungsstandards unter anderem verschiedene Leitideen ausgewiesen, in denen sich die mathematischen Kompetenzen zeigen sollen (Auszug aus der Modellierung der Kompetenzen Deutsch-Lesen und Mathematik in der Abbildung 3.1, Darstellung aus dem Kompetenzbrowser[2]). Um die zu messende Fähigkeit durch mehrere Aufgaben zu operationalisieren, müssen bei der Aufgabenauswahl alle Facetten in angemessener Weise berücksichtigt werden.

Für die fünf Leitideen der Mathematik gehen Kühn und Drüke-Noe (2013) offensichtlich davon aus, dass eine Gleichverteilung angemessen ist, wenn sie erwarten, dass sich

> „die Zusammenstellung [...] durch eine inhaltliche Reichhaltigkeit sowie eine hinreichende Breite der geforderten Kompetenzen *ohne eine einseitige Bevorzugung* bestimmter Kompetenzen und Komplexitätsgrade auszeichnen ... “ (ebenda, S. 923, Hervorhebung durch den Autor)

soll. Die Autorinnen geben dafür allerdings keine Begründung an. Sie ergänzen allerdings in den weiteren Ausführungen den noch zuvor ausgeführten Aspekt, wenn sie der angemessenen Repräsentation aller inhaltlichen Facetten jene von unterschiedlichen Komplexitätsgraden hinzufügen. Warum aber soll eine Gleichverteilung der Teilkompetenzen eine angemessene Repräsentanz darstellen?

Im November 1999 fand an der medizinischen Fakultät der erste Progress-Test statt und

[2] www.kompetenzbrowser.de

bezieht seitdem jährlich sämtliche Studierende der Medizin ein (Mertens et al., 2000; Charité - Universitätsmedizin Berlin, 2021). Die Entwicklung dieses Tests folgte internationalen Vorbildern. Er besteht aktuell aus 200 Fragen im Multiple-Choice-Format und stellt eine formative Prüfungsform dar, bei der Studierenden der Medizin aller Studiensemester eine jährlich fortgeschriebene Rückmeldung über ihre summarisch erworbenen Fähigkeiten erhalten. Für die Medizin wurde die Frage der inhaltlichen Repräsentanz ihrer spezifischen Fachgebiete über einen sogenannten Blueprint beantwortet. Mit dem *Masterplan Medizinstudium 2020* geht der darin noch als *Blueprint* bezeichnete Gegenstandskatalog in einem *Nationalen Kompetenzbasiertem Lernzielkatalog* (NKLM) (Medizinischer Fakultätentag, 2021) auf. In einem Tagungsbericht des außerordentlichen Medizinischen Fakultätentages 2002 in Mainz referiert Neuser (2002) zur Entwicklung solcher Gegenstandskataloge im internationalen Vergleich und stellt fest:

> „Die zusammenfassende Darstellung der Bemühungen, Prüfungsinhalte zu spezifizieren, muß zunächst ergeben, daß alle Kataloge ein das medizinische Wissen in seiner Gesamtheit umfassendes Konvolut ausmachen. In keinem Land der Welt ist nach unserer Kenntnis ein restringierter Ausschnitt medizinischen Wissens für die Ausübung des Arztberufes hinreichend. Differenzierte Kataloge beschränken sich nicht auf die Auflistung von Prüfungsgegenständen, sondern nehmen Gewichtungen nach der Bedeutung des Gegenstandes für den Arztberuf und nach den Kompetenzanforderungen in Bezug auf den jeweiligen Ausbildungsabschnitt vor.“ (ebenda S.5)

Für den medizinischen Bereich bilden beispielsweise die Häufigkeit der erteilten Diagnosen nach ICD-10 und deren Zuordnung zu medizinischen Fächern oder Organsystemen eine Basis für die Gewichtung der Prüfungsfragen im Progress-Test.

Für die Kompetenzen der Domänen der Bildungsstandards finden sich solcherlei Gewichtungen nicht. Wenngleich es schwierig sein mag ein Äquivalent zu finden, welches beispielsweise die Praxis der mathematischen Leitideen gleichermaßen abzubilden in der Lage ist, kann die unbegründete Annahme einer Gleichverteilung wenig zufriedenstellen. Zudem werden solche Gewichtungen tatsächlich vorgenommen, wenn im Rahmen des Designs der Pilotierungen die fünf Leitideen, der einfachen Organisation wegen, auf vier Module aufgeteilt werden.

> „Die entwickelten Aufgaben wurden auf vier Testhefte mit je vier Blöcken verteilt, wobei die Bearbeitungszeit jedes Blockes ungefähr 20 Minuten betrug. Die meisten Blöcke enthielten vornehmlich Items zu einer Leitidee. Block 4 bildete eine

> Ausnahme und enthielt etwa gleich viele Items zu den Leitideen 2 (Messen) und 3 (Raum und Form).“ (Aneis et al., 2018, S.7)

Bei den Vergleichsarbeiten in der Jahrgangsstufe 3 (VERA-3) setzten sich bis zur Durchführung 2019 die Tests für das Fach Mathematik aus zwei Aufgabenblöcken zusammen, welche jeweils Aufgaben aus genau zwei der fünf Kompetenzbereiche enthielten. Hier stellt sich die Frage, ob das Ergebnis eine Interpretation als globale Mathematik-Fähigkeit erlaubt. Im Rahmen der Modularisierung wurde auch für VERA-3 ab dem Jahr 2020 ein Basismodul ausgeliefert, welches Aufgaben aus allen fünf Kompetenzbereichen enthielt[3]. Die Pilotierung weist für die Grundschule hier eine Blockung aus, welche wieder eine Gleichgewichtung aller fünf Kompetenzbereiche unterstellt (Kohrt et al., 2020).

Der hier nur kurz diskutierte, oft aber ausgesparte Aspekt der Repräsentation aller Facetten einer Fähigkeit ist für die Interpretation einer Messung relevant, insbesondere aber, wenn zwei Messungen mit unterschiedlichen Instrumenten, also zwei ggf. unterschiedlichen Repräsentationen, aufeinander bezogen interpretiert werden sollen.

3.3. Rasch-Skalierung für dichotome Items

Im Zentrum der Abbildung einer Fähigkeit auf einer quantitativen Skala steht die Relation der Personenfähigkeit θ zur Wahrscheinlichkeit P einer korrekten Antwort $X = 1$ auf eine Aufgabe X, die als Funktion der Itemcharakteristik (Item Characteristic Curve, ICC) $f(\theta)$ beschrieben wird.

$$P(X = 1) = f(\theta)$$

Eine einfache Vorstellung (äquivalent dem Hochspringer von vorher) ist die, einer stufenförmigen Itemcharakteristik (Abbildung 3.2, links oben). Dabei lösen Personen mit einer Fähigkeit $\theta < \theta_x$ die Aufgabe nicht korrekt, Personen mit einer Fähigkeit von θ_x oder größer lösen die Aufgabe korrekt (blaue Funktion). Diese ICC weist damit am Punkt θ_x einen unsteten Verlauf auf.

$$P(X = 1) = \begin{cases} 0, & \text{wenn } \theta < \theta_x \\ 1, & \text{wenn } \theta \geq \theta_x. \end{cases}$$

[3] Für die Sekundarstufe I werden für das Fach Mathematik die fünf inhaltlichen Kompetenzen als *Leitideen* beschrieben. Diese unterscheiden sich von den für die Grundschule beschriebenen *Kompetenzbereichen* nur partiell. Zu den Unterschieden der Umsetzung der Bildungsstandards in den Rahmenplänen für Berlin/-Brandenburg gehört beispielsweise, dass genau diese Unterschiede zu Gunsten einer für alle Jahrgangsstufen von der ersten bis zur zehnten identischen Festlegung angepasst wurden.

Die Schwierigkeit einer Aufgabe σ wird hier als der Wert der Unstetigkeit auf der Skala der Fähigkeiten definiert $\sigma = \theta_x$. Eine ICC beschreibt für jeden Fähigkeitswert die Wahrscheinlichkeit für die richtige Lösung der konkreten Aufgabe. Die inverse Funktion stellt die Wahrscheinlichkeit dafür dar, dass die Aufgabe nicht richtig gelöst wird.

In der Abbildung 3.2 ist auf der linken Seite oben die ICC für eine Aufgabe mit der Schwierigkeit $\sigma_1 = -1$ dargestellt (blaue Funktion). Es ist die Wahrscheinlichkeit für die korrekte Lösung dieser Aufgabe und definitionsgemäß lösen Personen mit einer Fähigkeit die kleiner als -1 ist, diese Aufgabe nie ($P = 0$) und solche, mit einer Fähigkeit von -1 oder größer, immer ($P = 0$). Für eine zweite Aufgabe mit der größeren Schwierigkeit $\sigma_2 = 0,5$, ist mit der roten Funktion darunter die inverse ICC dargestellt, welche die Wahrscheinlichkeit dafür repräsentiert, dass eine Person diese schwierigere Aufgabe nicht korrekt löst. Für Personenfähigkeiten von unter 0,5 ist es sicher, dass die Aufgabe nicht gelöst wird; für Personen mit Fähigkeiten die gleich 0,5 oder größer sind, ist die Wahrscheinlichkeit gleich Null, die Aufgabe nicht zu lösen. Stellt man sich eine Person vor, welche die leichte Aufgabe mit der Schwierigkeit $\sigma_1 = -1$ korrekt löst, die schwierige Aufgabe mit $\sigma_2 = 0,5$ nicht, dann gelten für diese Person genau die beiden dargestellten Wahrscheinlichkeitsfunktionen. Betrachtet man das Lösen der zwei Aufgaben bis auf die Tatsache, dass sie die gleiche Fähigkeit beschreiben, als zwei voneinander unabhängige Ereignisse, dann ergibt sich die Wahrscheinlichkeit für das gleichzeitige Eintreten beider Ereignisse (leichte Aufgabe gelöst und schwere Aufgabe nicht gelöst) als multiplikative Verknüpfung[4]. Im Beispiel für die richtig gelöste leichte Aufgabe X_1 und die nicht korrekt gelöste schwierige Aufgabe X_2 ergibt sich daher die grüne Funktion in der Abbildung 3.2 (links unten)

$$P(X_1 = 1, X_2 = 0) = \begin{cases} 0, & \text{für } \theta < -1 \\ 1, & \text{für } -1 \leq \theta \leq 0,5 \\ 0, & \text{für } \theta > 0,5. \end{cases}$$

Die Fähigkeit dieser Person liegt somit zwischen der Schwierigkeit der richtig gelösten leichten Aufgabe X_1 und der Schwierigkeit der nicht richtig gelösten Aufgabe X_2, also zwischen -1 und 0,5.

Für die Bearbeitung von zwei Aufgaben durch eine Person sind vier Ergebnisse möglich.

[4]Die Wahrscheinlichkeit für das gleichzeitige auftreten zweier Ereignisse A und B, also $P(A \cap B)$ ist genau dann gleich der Multiplikation der Einzelwahrscheinlichkeiten $P(A) \cdot P(B)$, wenn die zwei Ereignisse A und B stochastisch unabhängig sind.

- Die Person löst die erste Aufgaben korrekt, die zweite schwierigere nicht. Wie dem obigen Beispiel zu entnehmen ist, ist für jede Fähigkeit zwischen den zwei Aufgabenschwierigkeiten dieses Antwortmuster eine modellkonforme Erklärung.

- Die Person löst beide Aufgaben korrekt. Damit kann für die Fähigkeit lediglich eine untere Schranke festgestellt werden. Die Fähigkeit der Person ist mindestens $\theta = \sigma_2$, kann aber auch beliebig darüber liegen. Die Funktion der bedingten Wahrscheinlichkeit wird auf der positiven Achse ab σ_2 immer bei 1 liegen. Modellkonform sind damit alle Personenfähigkeiten $\theta \geq \sigma_2$.

- Die Person löst beide Aufgaben nicht korrekt. Äquivalent ergibt sich hier nur eine obere Schranke. Die Fähigkeit der Person ist dann mit $\theta < \sigma_1$ modellkonform beschrieben.

- Die Person löst die zweite schwierigere Aufgaben korrekt, die erste allerdings nicht. Die Multiplikation der zwei ICCs ergibt kein Maximum, sondern ist überall Null, denn dieses Antwortverhalten ist nicht modellkonform. Aus dem Antwortverhalten kann keine Aussage zur Fähigkeit getroffen werden.

In gleicher Form kann das Antwortverhalten mit verschiedenen Formen von ICCs modelliert werden. In der Mitte der Abbildung 3.2 ist eine komplexere Form einer ICC dargestellt. Diese weist für die Aufgabe X_1 im Abschnitt der Fähigkeitsskala zwischen $\theta_{xu} = -2$ und $\theta_{xo} = 0$ einen linear ansteigenden Verlauf auf.

$$P(X = 1) = \begin{cases} 0, & \text{wenn } \theta < \theta_{xu} \\ \frac{\theta - \theta_{xu}}{(\theta_{xo} - \theta_{xu})}, & \text{wenn } \theta_{xu} \leq \theta \leq \theta_{xo} \\ 1, & \text{wenn } \theta > \theta_{xo}. \end{cases}$$

Die erste Aufgabe (blaue Funktion) wird bis zur Fähigkeit $\theta_{xu} = -2$ mit Sicherheit nicht korrekt und ab der größeren Fähigkeit $\theta_{xo} = 0$ sicher korrekt gelöst. Zwischen diesen zwei Punkten wird mit einer linearen Funktion eine Wahrscheinlichkeit angegeben, mit der die Aufgabe korrekt gelöst wird und die mit zunehmender Fähigkeit steigt. Mit dieser Formulierung einer ICC wird das Modell echt probabilistisch. Es gibt einen Bereich, in dem die Wahrscheinlichkeit der korrekten Lösung zwischen 0 und 1 liegt. Es kann demnach modellkonform sein, dass eine von zwei Personen mit identischer Fähigkeit eine konkrete Aufgabe richtig löst und die zweite Person nicht. Die Schwierigkeit einer Frage kann hier verschieden definiert werden, üblich wird der Punkt auf der Skala der Fähigkeit angegeben, bei

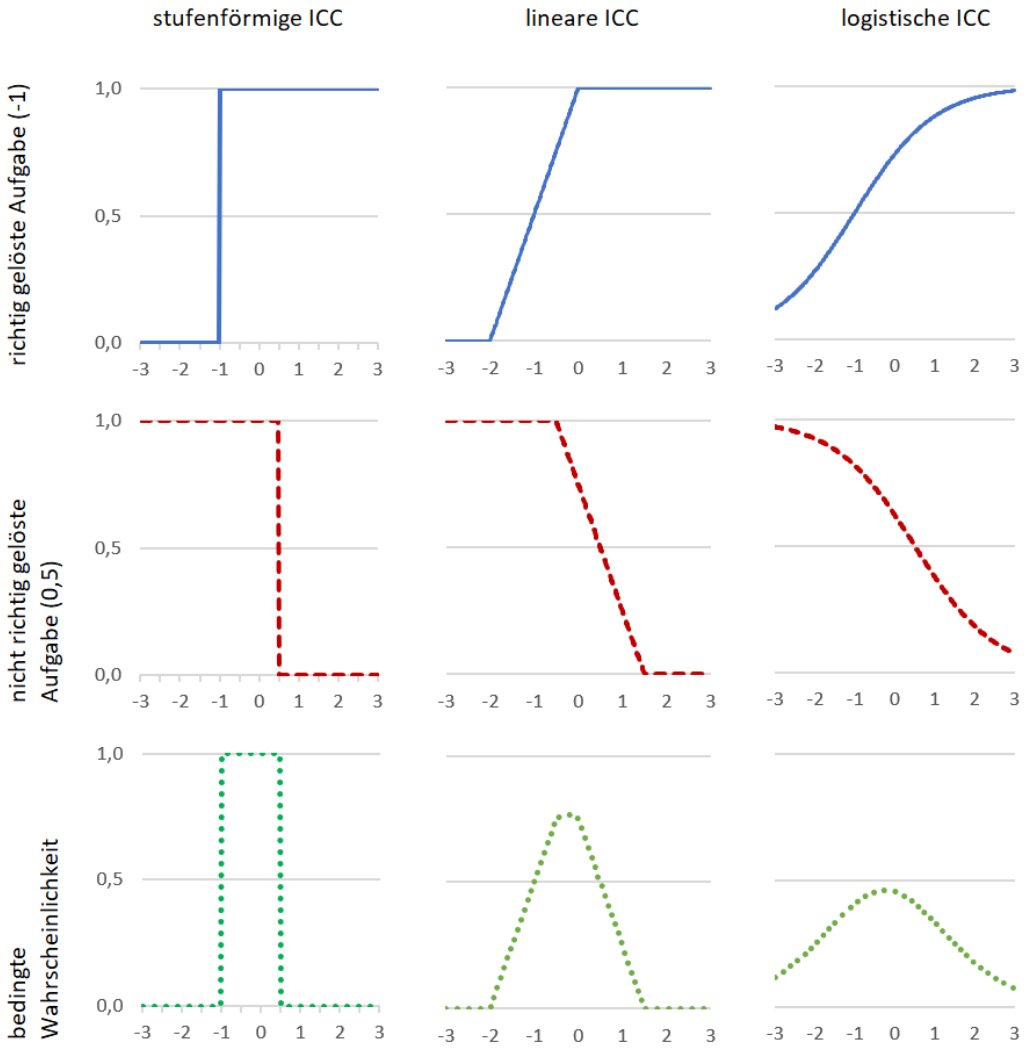


Abbildung 3.2.: Die Schätzung der Fähigkeit mit zwei Aufgaben bei verschiedenen Modellen von Itemfunktionen

dem die ICC eine Wahrscheinlichkeit von 0,5 aufweist, im vorliegenden Fall also der Punkt $(\theta_{xo} - \theta_{xu})/2 = -1$. Für eine Person, welche die erste Aufgabe richtig und die zweite nicht richtig löst, ist wieder die bedingte Wahrscheinlichkeit als grüne Funktion abgetragen[5]. Diese weist an der Fähigkeitsstelle von -0,25 ein Maximum auf, also genau zwischen den zwei Aufgabenschwierigkeiten von -1 und 0,5. Will man die Fähigkeit einer Person angeben, welche die erste Aufgabe korrekt löst und die zweite nicht, dann ist der Punkt des Maximums der bedingten Wahrscheinlichkeit die beste Schätzung. Auch bei dieser Modellierung sind einzelne Antwortmuster konstruierbar, die als nicht modellkonform gelten müssen und für die keine Aussage über die Personenfähigkeit getroffen werden können[6].

[5] Die zwei Aufgaben haben wieder die Schwierigkeiten von $\theta_{xu} = -1$ und $\theta_{xo} = 0.5$.

[6] Dies ist immer dann der Fall, wenn über das gesamte Fähigkeitsspektrum immer mindestens eine der sich durch die Antworten ergebenden Itemfunktionen Null ist, denn damit ergibt die multiplikative Verknüpfung

Auch das Modell von Georg Rasch (1960) modelliert die Abhängigkeit der Lösungswahrscheinlichkeit einer Aufgabe von der Fähigkeit als Wahrscheinlichkeitsfunktion, die hier aber einen über den gesamten Fähigkeitsbereich stetigen Verlauf hat (siehe Abbildung 3.2, blaue Funktion rechts oben), was den mathematischen Umgang vereinfacht.

$$P(X = 1) = \frac{e^{\theta-\sigma}}{1 + e^{\theta-\sigma}} \tag{3.1}$$

Die inverse ICC, also die Wahrscheinlichkeit dafür, dass eine Aufgabe nicht gelöst wird lässt sich mathematisch darstellen als

$$P(X = 0) = \frac{1}{1 + e^{\theta-\sigma}}$$

Die Fähigkeit θ überstreicht wie zuvor auf der Abszisse ein theoretisches Intervall von minus bis plus unendlich, während die Lösungswahrscheinlichkeit $P(X = 1)$ auf der Ordinate im begrenzten Intervall von 0 (Aufgabe wird sicher nicht gelöst) bis 1 (Aufgabe wird sicher gelöst) abgetragen wird. Die Modellierung des Zusammenhangs ist trivial: Mit zunehmender Fähigkeit steigt die Lösungswahrscheinlichkeit streng monoton an. Für sehr geringe Fähigkeiten nähert sich die Lösungswahrscheinlichkeit 0 und für sehr hohe Fähigkeiten 1, wobei diese Werte aber nie erreicht werden.

Untersucht man die Funktion der ICC mathematisch, findet man das Maximum der Steigung der ICC an der Nullstelle der zweiten Ableitung. Diese lautet mit $x = \theta - \sigma$

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\left(\frac{e^x}{1 + e^x}\right)\right) = -\frac{e^x(e^x - 1)}{(1 + e^x)^3} \tag{3.2}$$

und wird bei $x = 0$ und damit $e^0 = 1$ im Zähler und insgesamt gleich Null, also an der Stelle, wo die Fähigkeit θ gleich der Aufgabenschwierigkeit σ ist. An diesem Wendepunkt der ICC ist die Lösungswahrscheinlichkeit genau 0,5, ein Punkt, der auch im Rasch-Modell die Definition für die Schwierigkeit einer Aufgabe darstellt. Der Verlauf der inversen ICC für die nicht korrekt gelöste schwierige Aufgabe X_2 ist in Abbildung 3.2 wieder rot dargestellt. Als bedingte Wahrscheinlichkeit für die gelöste einfache und ungelöste schwierige Aufgabe ergibt sich wieder die grüne Kurve als multiplikative Verknüpfung der Wahrscheinlichkeiten für die richtig gelöste und die nicht richtig gelöste Aufgabe

$$P(X_1 = 1, X_2 = 0) = P(X_1 = 1) \cdot P(X_2 = 0) = \frac{e^{\theta-\sigma_1}}{1 + e^{\theta-\sigma_1}} \cdot \frac{1}{1 + e^{\theta-\sigma_2}} \tag{3.3}$$

für jeden Punkt der Skala Null.

Für die beste Schätzung der Personenfähigkeit θ muss das Maximum dieser Funktion, also die Nullstelle der ersten Ableitung gefunden werden. Da keine geschlossene Lösung existiert, müssen numerische Verfahren zu Hilfe gezogen werden. Dieses Beispiel mit zwei Aufgaben, lässt sich äquivalent auf beliebig viele Aufgaben erweitern. Die beschriebene Bestimmung von Personenfähigkeiten geht davon aus, dass die Schwierigkeiten der Aufgaben σ_i bekannt sind. Für den hier betrachteten Kontext der Vergleichsarbeiten ist dies gegeben. Zudem ist keine dieser Wahrscheinlichkeitsfunktionen jemals echt Null. In der Folge ergibt sich in jedem Fall ein Maximum der bedingten Wahrscheinlichkeiten, auch wenn dieses ggf. sehr klein ist. Damit führt (fast) jedes Antwortmuster zu einer Fähigkeitsschätzung. Keines ist per Definition unmöglich also nicht modellkonform, höchstens eben unwahrscheinlich. Nicht modellkonform sind lediglich in einer (Teil-)Population überhäufig vorkommende, unwahrscheinliche Antwortmuster.

Als Folge der beschriebenen Modellierung der ICC über eine logistische Funktion, schneiden sich zwei ICCs von Aufgaben verschiedener Schwierigkeit im Rasch-Modell nie. Die ICCs von Aufgaben unterschiedlicher Schwierigkeiten ergeben sich quasi als Verschiebung der ICC für eine schwierigere Aufgabe nach rechts und für eine leichtere Aufgabe nach links.

Dass sich zwei ICCs nicht schneiden, bedeutet, dass für zwei Personen unabhängig von ihrer Fähigkeit immer die gleiche von zwei Aufgaben die Schwierigere ist. Hier manifestiert sich die Grundannahme der eindimensional modellierten Fähigkeit und es verwundert kaum, wenn Hans Spada zu dem Schluss kommt: „Die Realität ist deutlich komplexer als das Rasch-Modell.“(Rost, Jürgen, 2005, ab 1:23:10). Auch Kolen und Brennan (2014) schreiben:

> „However, that assumed form of the relationship between ability and the probability of a correct response is choose primarily for reasons of mathematical tractability. No reason exists for this relationship to hold, precisely, for actual test items.“(Kolen & Brennan, 2014, S.160)

Der Tatsache der Simplizität der gewählten Modellfunktion wird oft mit der Verwendung komplexerer Modellfunktionen begegnet. Auch jede andere Funktion wird aber die Realität nur mit begrenzter Präzision abbilden können. Komplexere Modellfunktionen beziehen weitere zu bestimmende Parameter ein. Es ist zu prüfen, ob komplexere Modelle gerechtfertigt sind. Oft wird diese Prüfung und Entscheidung auf der Basis mathematischer Parameter getroffen, welche eine Modellgüte bewerten, also überprüfen, wie gut das gewählte Modell unter Einbezug der Zahl an Freiheitsgraden in der Lage ist, die realen Daten zu erklären. Insbesondere im Rahmen von Vergleichsarbeiten, deren Ergebnisse im praktischen Kontext interpretiert wer-

den, muss sich ein Modell aber auch im Rahmen dieser praktischen Verwendung als plausibel erweisen.

Die ICC einer Aufgabe im Rasch-Modell wird nur durch einen einzigen Parameter bestimmt, die Schwierigkeit. Ein weiterer, schon aus der klassischen Testtheorie bekannter Parameter einer ICC ist die Trennschärfe. Die Trennschärfe δ ist im Rasch-Modell gleich 1 und insofern stellt das sogenannte 2PL-Modell (two-parameter-logistic model) eine Verallgemeinerung des Rasch-Modells dar. Die ICC einer Aufgabe mit höherer Trennschärfe δ weist eine größere Steigung auf.

$$P(X = 1) = \frac{e^{\delta(\theta-\sigma)}}{1 + e^{\delta(\theta-\sigma)}} \tag{3.4}$$

Im Rahmen von PISA wurde anfangs das Rasch-Modell zu Grunde gelegt (OECD, 2003), ab 2015 wechselte man zum 2PL-Modell (OECD, 2017), dass auch als Birnbaum-Modell bezeichnet wird. Die Steigung ist, wie zuvor schon beschrieben, natürlich nicht an allen Punkten einer ICC identisch. Sie wächst von Null kommend bis zum Wendepunkt bei einer Lösungswahrscheinlichkeit von 0,5 an, hat dort ihr Maximum und sinkt dann wieder kontinuierlich. Wenn also dem Parameter der Trennschärfe in der Folge ein Wert zugewiesen wird, meint dies immer den im Wendepunkt erreichten Maximalwert der Steigung. Im Rasch-Modell ist der Verlauf aller ICCs identisch und somit auch die Ableitung jeder ICC. Variiert man die Steigung hat dies mehrere Implikationen. Zuerst ergibt sich durch die Variation von zwei Parametern eine bessere Möglichkeit, eine Passung einer solchen ICC an vorhandene Daten herzustellen. Konkret wird die Passung mindestens identisch ausfallen, im Allgemeinen aber besser werden und niemals schlechter. Dieser Vorteil wird aber durch zusätzliche Parameter „erkauft“. Was bedeutet dieser neue Parameter aber abseits einer mathematischen Betrachtung praktisch? Er beschreibt die Trennschäfte einer Aufgabe, also die Fähigkeit einer Aufgabe, zwischen Personen zu differenzieren, die diese Aufgabe lösen und jenen, die sie nicht lösen. Die Stufenfunktion kann man als ICC einer Aufgabe mit extremer - tatsächlich unendlicher - Trennschäfte ansehen[7]. Dass Aufgaben verschiedene Trennschäften aufweisen, findet sich tatsächlich auch in der Praxis. Modelliert man eine Fähigkeit im Rahmen eines Tests mit dem 2PL-Modell, ist damit die Möglichkeit gegeben, dass sich zwei ICCs schneiden. In der Konsequenz gilt nicht mehr, dass für zwei Personen unabhängig von ihrer Fähigkeit immer die gleiche von zwei Aufgaben die schwierigere ist. Zudem weist das Rasch-Modell einige mathematische Einfachheiten auf, die mit dem Einbezug dieses zweiten Parameters verloren

[7] Wieder ist die Trennschärfe die Steigung in dem einen Wendepunkt gemeint, der hier an der Position der Stufe zu verorten ist.

gehen. Das 3PL-Modell führt mit einem sogenannten Rate-Parameter γ einen weiteren Parameter in das Modell ein, wobei die Kennzeichnung *3PL* suggeriert, dass dieser Parameter zwingend *nach* dem Trennschärfe-Parameter hinzukommt. Tatsächlich können Trennschäfte und variabler Rate-Parameter unabhängig voneinander Bestandteil der Modellierung einer ICC sein.

$$P(X = 1) = \gamma + (1 - \gamma)\frac{e^{\delta(\theta-\sigma)}}{1 + e^{\delta(\theta-\sigma)}} \tag{3.5}$$

Für die Genese dieses dritten Parameters ist der Einsatz von Multiple-Choice-Aufgaben ein praktischer Ausgangspunkt. Soll beispielsweise aus vier Alternativen die eine korrekte ausgewählt werden, muss bei genügend großer Zahl von Personen angenommen werden, dass die Lösungswahrscheinlichkeit bei mindestens 25% liegt, da selbst eine vollständige Zufallsauswahl zu diesem Ergebnis führen würde. Dass mit der ICC die Lösungswahrscheinlichkeit bei einer Fähigkeit von minus unendlich gegen Null strebt, ist dann nicht plausibel. Der dritte Parameter behebt dieses praktische Problem, in dem die ICC hier gegen den Rate-Parameter strebt. Grundsätzlich können die Aufgaben eines Tests durchaus unterschiedlich modelliert werden. So wird man dem Beispiel folgend bei Multiple-Choice-Aufgaben einen Rateparameter einbeziehen, bei anderen Aufgaben gerade nicht.

Im Rahmen der Vergleichsarbeiten wurde bis heute ausschließlich das Rasch-Modell verwendet und bis auf eine Aufgabe im ersten VERA-8-Mathematik-Test wurden auch sämtliche Aufgaben als dichotom, also nur als richtig vs. falsch kodiert. Deshalb beschäftigen sich die folgenden Untersuchungen auch lediglich mit dem Rasch-Modell und dichotomen Aufgaben. Darüber hinaus wird lediglich die Trennschäfte thematisiert.

3.4. Standard-Setting

In den vorangegangenen Abschnitten wurde eine angemessenen Auswahl von Aufgaben vorgenommen, Aufgaben, die einer an der fachlichen Beschreibung einer Fähigkeit bzw. Kompetenz orientierten Operationalisierung entstammen. Diese Aufgaben haben sich folgend gegenüber dem Rasch-Modell als konform erwiesen und so konnte eine Metrik für die Fähigkeit definiert werden. Nun sollen auf dieser Metrik fachinhaltlich beschreibbare Fähigkeitsbereiche gegeneinander abgrenzen werden. Verfahren, welche die Festlegung solcher als Cut-Scores bezeichneten Grenzwerte auf der Metrik generieren und damit eine Brücke von der psychometrischen Skala zur fachlichen Bewertung von Ergebnissen schlagen, werden als Standard-Setting (Pant et al., 2010) bezeichnet. Die Bookmark-Methode ist ein Verfahren, welches folgend beispiel-

gebend illustriert werden soll. Im Rahmen der Entwicklung von Kompetenzstufenmodellen für das Fach Englisch mit Bezug auf die Bildungsstandards der KMK wurde dieses Verfahren angewendet und ist von Tiffin-Richards et al. (2013) ausführlich dokumentiert.
Ausgangspunkt der Bookmark-Methode ist das für die entsprechende Fachdomäne vorliegende Kompetenzstufenmodell, auf dessen Basis das Instrument zur Messung der entsprechenden Domäne entwickelt worden ist und für das auch Ergebnisse aus einer Normierung an der in Frage stehenden Population vorliegen. Im Rahmen einer dem eigentlichen Standard-Setting vorgeschalteten Familisierungsphase versichert sich eine Gruppe ausgewiesener Experten*innen aus Fachdidaktiker*innen und Fachlehrkräften eines gemeinsamen Verständnisses der zugrundeliegenden Beschreibungen der Kompetenzstufen. Allen Teilnehmer*innen des Standard-Settings wird dann das nach empirischer Schwierigkeit geordnete Set von Aufgaben vorgelegt (ordered item booklet, OIB), wobei die Schwierigkeiten aus der Erhebung an der Normierungsstichprobe resultieren. Die Experten werden zuerst für sich allein um eine Einschätzung gebeten, im OIB beginnend bei der leichtesten Aufgabe jene erste zu finden, deren Lösung Anforderungen benötigt, die nicht allein mit jenen Fähigkeiten erfolgreich bearbeitet werden können, die dem Bereich der ersten Stufe zuzurechnen werden. Die üblich differenten Einschätzungen der Teilnehmer*innen werden vor dem Hintergrund der Kompetenzstufenbeschreibungen diskutiert, bis sich nach einer zweiten und dritten jeweils eigenständigen Einschätzung mit anschließender Diskussion im Allgemeinen ein Konsens einstellt. Dieses Prozedere wird für jede Stufengrenze wiederholt. Beim Einsatz der Bookmark-Methode im Rahmen der Bestimmung der Stufengrenzen für die Modelle, die auch den Vergleichsarbeiten zu Grunde liegen, wurde nach der zweiten und vor der dritten Runde eine Verteilung der Kompetenzstufen präsentiert, wie sie sich aus der gemittelten Einschätzung der Expert*innen ergeben würde. Es kann angenommen werden, und es entsprach sicher auch der Intention der das Standard-Setting Durchführenden, dass diese Intervention ggf. zu einer Korrektur der Einschätzung der Expert*innen geführt hat. Es ist zu diskutieren, inwieweit diese Intervention hilfreich ist. Sie führt sicherlich dazu, dass die durch die Stufengrenzen formulierten Anforderungen die Praxis nicht über- oder unterfordern. Andererseits bleibt die Frage, ob die fachdidaktische Einschätzung hierdurch unzulässig verzerrt wird.

Die so festgelegten Grenzen können dann in dieser Form für zukünftige Erhebungen als kriteriale Beschreibung mitgegeben werden, wie es beispielsweise im Rahmen der Österreichischen Standardüberprüfungen geschieht (Schreiner et al., 2020). Für die Mathematik der

Sekundarstufe sind hier 439, 517 und 690 als Grenzen in einer 500/100er Metrik[8] angegeben. Im Gegensatz dazu sind die Grenzen im 6-stufigen Modell der Bildungsstandards in Deutschland mit 355, 435, 515, 595 und 675 äquidistant festgelegt worden (Aneis et al., 2020). Die letztendliche Festlegung der Stufengrenzen erfolgte hierbei nach dem Standard-Setting in Absprachen mit der KMK, die innerhalb jeder Domäne gleiche Abstände forderte.

Das Ergebnis des Standard-Setting-Prozesses ist eine empirisch validierte Beschreibung messbarer Kompetenzstände, die im besten Fall konkrete Hinweise zur Förderung oder einer anschließenden, vertiefenden Diagnostik geben kann.

3.5. Testdesign

Im Abschnitt 3.2 wurde dargestellt, dass eine angemessene Repräsentation a) des zu messenden Konstrukts in seinen verschiedenen Facetten und b) der verschiedenen, den Fähigkeiten der Untersuchungsgruppe angepassten Aufgabenschwierigkeiten, den Einbezug einer größeren Zahl von Aufgaben verlangt. Entsprechende Test könnten dann allerdings unpraktikabel lang, die zeitliche und kognitive Belastung der Schülerinnen und Schüler zu hoch werden. Besteht, wie bei Large Scale Assessments, die Aufgabe eines Tests allerdings darin, für eine Grundgesamtheit einen Überblick über die Lernergebnisse zu erlangen, nutzt man bestimmte Testdesigns, bei denen nicht jede Schülerin und jeder Schüler alle Aufgaben, sondern nur eine Auswahl davon bearbeitet. So, wie man auf der Ebene der zu untersuchenden Personen eine Grundgesamtheit definiert und aus dieser eine repräsentative Stichprobe zieht, kann man das äquivalent auch für die Aufgaben tun. Die Grundgesamtheit aller Aufgaben, die eine zu messende Kompetenz darstellen, ist quasi unendlich groß, weshalb man oft vom Aufgabenuniversum spricht. Grundsätzlich besteht nun die Anforderung, aus diesem Aufgabenuniversum mehrere Stichproben zu ziehen, um diese auf die Personen-Stichproben zu verteilen. Die geplante Zuordnung von Personenstichproben zu Aufgabenstichproben wird als Testdesign bezeichnet, an das verschiedene Anforderungen gestellt werden. Das Ziel ist, dass die Messung der Kompetenz durch die Auswahl von Items aber auch deren Positionen im Testheft unbeeinflusst bleibt. Da bei Large-Scale-Assessments wie dem Bildungstrend üblicher Weise eine erhebliche Anzahl von Aufgaben eingesetzt wird, fasst man heute Aufgaben zu Aufgabenblöcken zusammen, für die dann die folgenden Anforderungen an das Testdesign erfüllt werden sollen (Becker et al., 2019, S.414):

[8] Die Kennzeichnung 500/100 meint, dass die Metrik an der Zielpopulation auf einen Mittelwert von 500 und eine Standardabweichung von 100 normiert ist. Eine solche Normierung ist sicher auch wegen der Verwendung bei PISA weit verbreitet.

- Die Zahl der Testhefte, in denen ein Aufgabenblock auftritt ist für alle Aufgabenblöcke gleich.
- Die Zahl der Aufgabenblöcke in einem Testheft ist identisch.
- Die Zahl der Paarungen von Aufgabenblöcken über alle Testhefte hinweg ist gleich.

Durch die Aggregation der Ergebnisse über die verschiedenen Testheftvariationen und Personenstichproben, kann dann eine Aussage getroffen werden, die sich auf eine angemessen vollständige Repräsentation der Kompetenz stützt. Balanciert ist ein Design, wenn jeder Aufgabenblock an jeder möglichen Position einmal vorkommt. Im besten Fall versuchte man sicherzustellen, dass jeder Aufgabenblock mit jedem anderen Aufgabenblock einmal in einem Testheft zusammen auftritt. Ein solches Design wird dann vollständig genannt. Der Aufwand für ein vollständiges und balanciertes Design kann dabei recht groß werden. So wurden für den Bildungstrend 2018 (Becker et al., 2019, S.415) allein für Mathematik ohne Berücksichtigung von Kindern mit sonderpädagogischem Förderbedarf 41 Aufgabenblöcke in die Testung einbezogen, wobei jeder Person genau sechs Blöcke vorgelegt wurden.

Die Abbildung 3.3 stellt links schematisch dar, wie alle Personen einer Stichprobe sämtliche verschiedenen Aufgaben bearbeiten. Um die Bearbeitungszeit, respektive die Zahl der zu bearbeitenden Aufgaben, zu reduzieren, wird die Zahl der Personen im mittleren Testdesign verdoppelt, so dass jede Person nur noch die Hälfte der Aufgaben bearbeiten muss. Die Zahl der Aufgabenbearbeitungen insgesamt bleibt dabei unverändert. Im Testdesign rechts sind sowohl Aufgaben wie Personen zu je vier Blöcken zusammengefasst worden. Die Aufgabenblöcke sind mit A, B, C und D bezeichnet. Im vorliegenden Beispiel werden jeder Person genau zwei Aufgabenblöcke zur Bearbeitung zugewiesen. Diese sind so verteilt, dass den oben beschriebenen Anforderungen genügt wird: Es gibt vier verschiedene Testhefte, welche in den vier Personengruppen repräsentiert sind und jeder Aufgabenblock kommt genau zwei Mal in allen Testheften vor. Dabei enthält jedes Testheft genau zwei Aufgabenblöcke. Überdies existiert jedes Paar von Aufgabenblöcken genau einmal.

Im Allgemeinen muss aber auf ein Balanced-Incomplete-Block-Design zurückgegriffen werden, in dem nicht jede mögliche Paarung von Aufgabenblöcken abgebildet wird. Schon für den oben beschriebenen Fall mit 41 Aufgabenblöcken und 6 Blöcken je Testheft ergibt sich eine erhebliche Anzahl an Varianten. Hier kommen Verfahren zum Einsatz, welche eine bestmögliche Verteilung der Aufgabenblöcke auf eine überschaubare Zahl an Testheftvarianten ermöglicht. Ausführliche Erörterungen dazu inkl. Hinweisen zu theoretischen Vorarbeiten finden sich beispielsweise bei Becker et al. (2019). Hier soll lediglich deutlich gemacht werden,

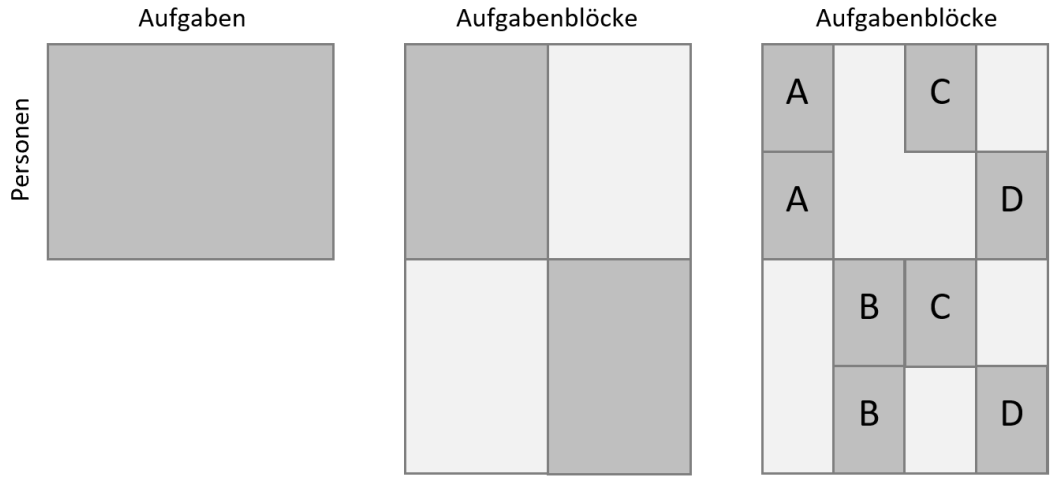


Abbildung 3.3.: Schematische Darstellung eines *balanced-complet-block*-Designs

dass bei den Erhebungen im Rahmen von Large-Scale-Assessments mit der Entwicklung und Umsetzung angemessener Testdesigns erheblicher Aufwand betrieben wird, um a) eine angemessene Repräsentation jeder Domäne sicherzustellen und b) diese einer repräsentativen Stichprobe von Schülerinnen und Schülern der Zielpopulation vorzulegen. Für die Stichprobenziehung von Schülerinnen und Schüler wird hier mit Mahler et al. (2019, S.110-114) nur beispielhaft auf ein Kapitels des Berichts zum Bildungstrend verwiesen.

3.6. Linking

Oft ist es nicht möglich oder nicht angezeigt, ein identisches Instrument für eine wiederholte Erhebung eines Merkmals zu nutzen. So wird beispielsweise für einen Schulabschluss als zertifizierende Prüfung verlangt, dass jedes Jahr neue Aufgaben zum Einsatz kommen. Allerdings werden damit zwei nachvollziehbare Forderungen verbunden:

- Die Repräsentation des zu messenden Merkmals muss äquivalent, zumindest aber angemessen ähnlich sein.

- Die Metriken beider Messungen sollen identisch sein, mindestens aber durch einen angebbaren Transfer aufeinander bezogen werden können.

Mit dem *Linking* soll sichergestellt werden, dass die Ergebnisse von Messungen aufeinander bezogen werden können, die mit unterschiedlichen Instrumenten das identische Konstrukt zu messen vorgeben. Für eine umfassende Beschreibung solcher Prozeduren sei hier auf das Standardwerk von Kolen und Brennan (2014) verwiesen. Folgend wird die Funktionsweise

nur skizzenhaft wiedergegeben und dabei auf die konkrete Anwendung im Rahmen der Skalierung mit IRT-Modellen in einem Design mit nicht-äquivalenten Gruppen mit gemeinsamen Aufgaben fokussiert. Dies ist ein weit verbreitetes Testdesign, welches dem rechten Schema in Abbildung 3.3 entspricht.

Bei Lee und Ban (2009) wird mit Bezug auf Kolen und Brennan (2014) *Linking* als jener Prozess beschrieben, mit dem an unterschiedlichen Personengruppen erhobene Daten auf einer gemeinsamen Skala verortet werden. Wie Hambleton et al. (1991) feststellen, ist bei bekannten Schwierigkeitsparametern der Aufgaben kein explizites Linking notwendig. Wenn einer Person Aufgaben mit bekannten Schwierigkeitsparametern vorgelegt werden, dann ergibt sich unabhängig von der Aufgabenauswahl immer ein identischer Schätzer für die Fähigkeit. Eine ungünstige Wahl von Aufgaben, wenn zum Beispiel überwiegend unpassend leichte oder schwierige Aufgaben verwendet werden, führt der Item-Response-Theorie folgend lediglich zu größeren Messfehlern.

Ausgangspunkt der Definition einer Metrik ist die Normierung. Dabei werden zuvor in einer Pilotierung in ihrer Funktion überprüfte Aufgaben einer repräsentativen Stichprobe der Grundgesamtheit vorgelegt. Bei der Normierung liegt der Fokus auf der Festlegung der Schwierigkeitsparameter der eingesetzten Aufgaben, die auf die Verteilung der Fähigkeitsparameter der Grundgesamtheit, repräsentiert durch die Normierungsstichprobe, bezogen wird. Bei der Skalierung von unbekannten Aufgaben muss ein Parameter fixiert werden, weil die zu konstituierende Skala bezüglich einer linearen Transformation unbestimmt ist. Das bedeutet, dass eine Kompetenzskala keinen natürlichen Fixpunkt hat und im Rahmen der Definition einer Skala ein solcher gefunden werden muss. Üblich wird die mittlere Fähigkeit auf Null gesetzt. Hier wird ersichtlich, wie relevant die Ziehung der Stichprobe ist. Ist diese nicht repräsentativ ist der gewählte Fixpunkt inkorrekt. Bei PISA wie auch für die BiSta-Skalen der im Bildungstrend und bei den Vergleichsarbeiten gemessenen Kompetenzen wird durch eine geeignete Transformation der Mittelwert der Grundgesamtheit auf 500 festgelegt[9]. Die in der Normierung verwendeten Aufgaben sind über die Skalierung bezüglich ihrer Schwierigkeit festgelegt. Mit der Transformation der Fähigkeitsparameter auf die BiSta-Skala können auch die Aufgabenschwierigkeiten äquivalent transformiert werden.

Wenn die in der Normierung eingesetzten Aufgaben öffentlich bekannt, verbreitet und ggf. geübt werden, muss davon ausgegangen werden, dass die ermittelten Schwierigkeiten keinen

[9]Bei der Rasch-Skalierung wird mit der Festlegung des Mittelwerts auf Null auch die Standardabweichung auf 1 festgelegt. Durch eine entsprechende Transformation wird diese 0/1-Skala in eine 500/100-ter Skala überführt

langfristigen Bestand haben; die Aufgaben gelten dann als *verbrannt*. Nur, wenn Erhebungen unter streng kontrollierten Bedingungen durchgeführt werden, können die Aufgaben aus der Normierung eingesetzt werden. Zentral geschulte Testleitungen sorgen dabei für kontrolliert standardisierte Durchführungsbedingungen. Da für die Vergleichsarbeiten jedes Jahr neue Aufgaben benötigt werden, stellt sich das Problem, wie für neu entwickelte Aufgaben der äquivalente Schwierigkeitsparameter gefunden werden kann. Eine Möglichkeit ist es, an einer Stichprobe der Grundgesamtheit einige aus der Normierung stammende Aufgaben mit bekannter Schwierigkeit zusammen mit neuen Aufgaben unbekannter Schwierigkeit einzusetzen. In der Stichprobe werden die zwei Aufgabengruppen so miteinander verlinkt, dass die unbekannten Schwierigkeiten der neuen Aufgaben auf der Basis der bekannten Schwierigkeiten der Normierungsaufgaben geschätzt werden können. Die Items aus der Normierung werden dabei als Ankeritems bezeichnet. Solch ein Linking findet in einer speziellen Studie statt. Beim Einsatz der neuen Aufgaben können die abgeleiteten Schwierigkeiten wie solche aus der Normierung verwendet werden. Das Ergebnis eines Tests mit neuen Aufgaben kann dann auf der vormals mit den Normierungsitems konstituierten Metrik abgebildet werden.

Von verschiedenen Autoren (beispielsweise Hambleton et al., 1991, S.125; Trendtel et al., 2016, S.212) werden dabei Anforderungen an das Linking bzw. die zu verlinkenden Tests formuliert:

- Beide Tests müssen ein identisches Konstrukt messen.
- Beide Tests müssen eine vergleichbare Reliabilität aufweisen.
- Zwei Tests mit unterschiedlichem Schwierigkeitsgrad dürfen nicht miteinander verlinkt werden.
- Das Ergebnis eines Linkings soll unabhängig davon sein, ob das Testergebnis von Test A auf der Skala des Tests B abgebildet wird oder umgekehrt.
- Das Ergebnis des Linkings soll unabhängig von der gezogenen Stichprobe sein.
- Eine Person soll bei beiden Tests das gleiche Ergebnis zeitigen.

Für das Linking bedeutet das in der Konsequenz, dass einerseits die Ankeritems für das Konstrukt und andererseits die Personen-Stichprobe für die Grundgesamtheit repräsentativ sein müssen und das die Bedingungen der Testadministration konstant und die zu verlinkenden Instrumente bezüglich Inhalt und Schwierigkeit parallel gehalten werden müssen.

3.7. Umsetzung für VERA

Im Unterschied zur Erhebung beim Bildungstrend, wo mit Hilfe von schulexternen Testleitungen stabile, geschützte und kontrollierte Durchführungsbedingungen geschaffen werden und so auch bei der Testwiederholung das identische Instrument erneut eingesetzt werden kann, müssen für die Vergleichsarbeiten jedes Jahr neue Testaufgaben entwickelt werden. Diese sollen für die unterrichtliche Weiterarbeit genutzt werden. Das Test- und Begleitmaterial soll so auch zu einer Dissemination der zugrundeliegenden Kompetenzorientierung führen. Das steht einer Wiederverwendung eines Instruments entgegen. Damit aber die Ergebnisse der Messung mit jedem Instrument immer auf die identische Skala der Bildungsstandards bezogen werden können, müssen die neu entwickelten Aufgaben nicht nur pilotiert und normiert werden, sondern für die Aufgaben müssen im Rahmen eines Linkings auch die entsprechenden Schwierigkeitsparameter gefunden werden. Nur mit diesem stabilen Bezug der Messung auf die BiSta-Metrik, gelten für diese Messergebnisse auch die mit der Metrik verbundenen Kompetenzstufen. Damit können die Ergebnisse für jede Kompetenz gleichermaßen über die Jahre hinweg und auch mit den Ergebnissen des Bildungstrends verglichen werden. Mit dem Linking wird eine psychometrische Kongruenz herbeigeführten. Die inhaltliche Gleichheit zweier Tests (siehe Abschnitt Operationalisierung 3.2) muss allerdings darüber hinaus sichergestellt werden, bzw. ist schon für das Linking als Anforderung beschrieben.

Die Darstellung 3.4 bildet einige der Prozessschritte ab die notwendig sind, um die Testhefte der Vergleichsarbeiten an die BiSta-Metrik anzubinden. Die Aufgabenentwicklung, die Pilotierung der Aufgaben und die Auswahl der am besten Geeigneten sind vorgelagert und in diese Darstellung nicht einbezogen. Die Basis bildet die Normierung, an Hand derer die Metrik der Bildungsstandards und die Kompetenzstufen definiert werden. Zudem wird durch die Festlegung der Schwierigkeitsparameter für die Normierungsaufgaben ein Pool von Aufgaben geschaffen, der für zukünftige Erhebungen des Kompetenzstandes eingesetzt werden kann. Um eine stabile und von Fehlern möglichst gering beeinträchtigte Festlegung der Metrik zu bewerkstelligen, werden verschiedene Anforderungen definiert (Becker et al., 2019). So wird die Stichprobe für die Normierung in einem aufwändigen mehrstufigen Verfahren aus Schulen in ganz Deutschland realisiert. Das hier nur beispielhafte Feld der Mathematik wird in 31 Testheften mit Aufgaben repräsentiert, deren Bearbeitungszeit summarisch 620 Minuten in Anspruch nimmt[10].Für die Mathematik der Sekundarstufe werden allerdings auch

[10]Hierbei wurde der erste Ländervergleich für das Fach Mathematik im Jahr 2012 zu Grunde gelegt und dabei nur jene Aufgaben, welche von Schülerinnen und Schülern ohne sonderpädagogischen Förderbedarf bearbeitet werden sollten. Diese, sowie Aufgaben zur Verlinkung mit den Aufgaben der Naturwissenschaften

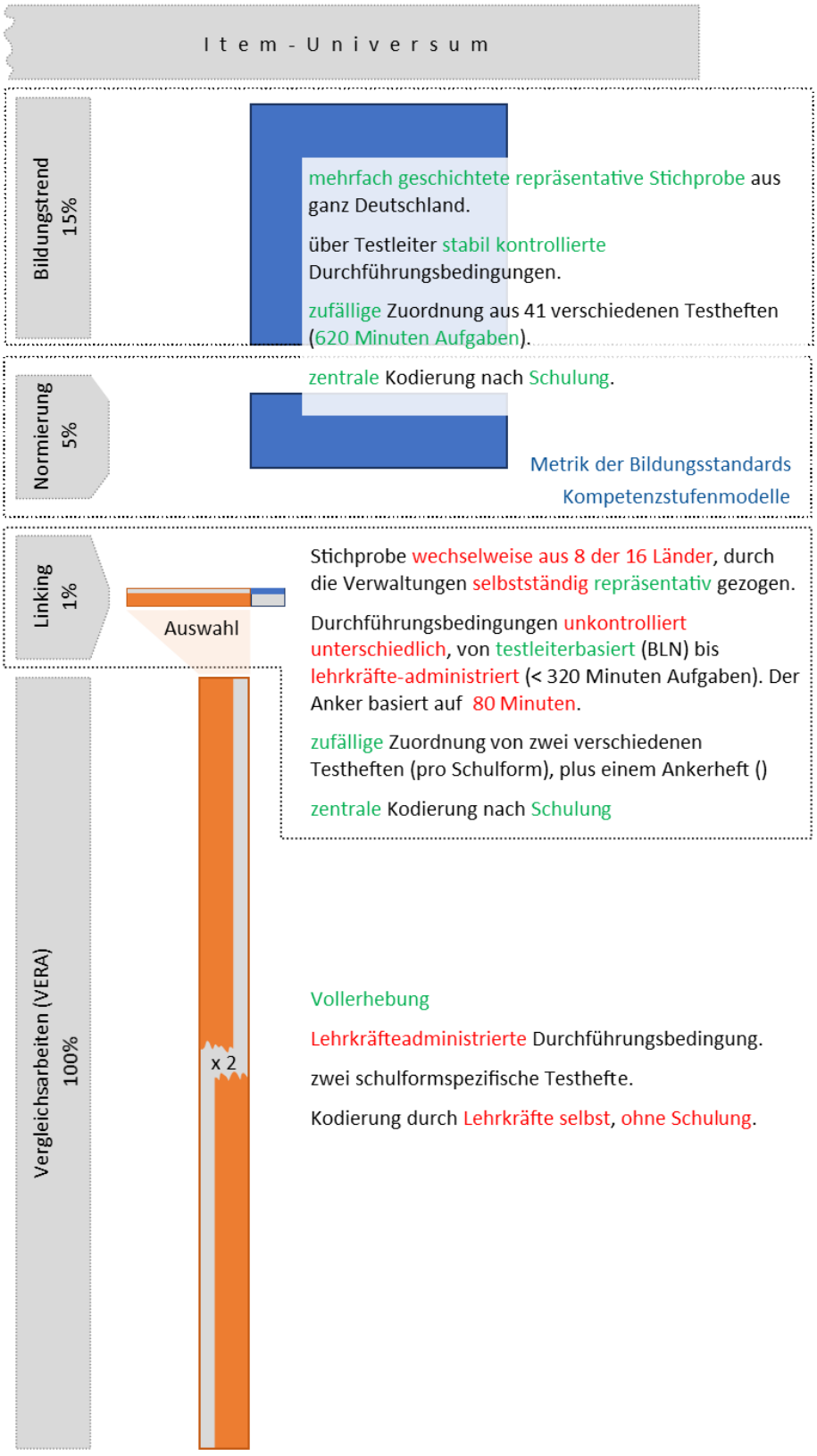


Abbildung 3.4.: Schematische Darstellung der Verlinkung des Bildungstrends sowie der Vergleichsarbeiten mit der Metrik der Bildungsstandards.

5 inhaltsbezogene Leitideen und 6 prozessbezogene Kompetenzen definiert, die sich zudem in drei verschiedenen Anforderungsbereichen spiegeln, so dass eine derart umfangreiche Operationalisierung gerechtfertigt erscheint. Der administrative Rahmen der Normierung sichert eine bestmögliche Stabilität der Messung ab.

Zur Durchführung des Bildungstrends wurde auf den identischen Aufgabenpool zurückgegriffen, eine für den Vergleich von Länderergebnissen auch für Teilgruppen notwendige größere Stichprobe der Grundgesamtheit gezogen und die Durchführungsbedingungen unverändert gelassen. Damit war sichergestellt, dass die Ergebnisse der Messungen auch ohne ein Linking auf der Skala liegen, die mit der Normierung definiert worden ist. Auch in den Folgeerhebungen des Bildungstrends wurden die Aufgaben nur teilweise durch Neuentwicklungen ersetzt und die Durchführung blieb unverändert.

Für die jährlich neuen Testhefte der Vergleichsarbeiten, müssen die Schwierigkeitsparameter der Aufgaben in einem Linking bestimmt werden. Hierzu wird das Linking als kontrollierte Erhebung umgesetzt, so dass Ankeraufgaben aus der Normierung eingesetzt werden können. Tatsächlich wird das Linking der Items mit der BiSta-Metrik und die Pilotierung der Aufgaben in einer gemeinsamen Studie durchgeführt. Um die Belastungen für die Länder und deren Schulen weiter zu reduzieren sowie die finanziellen Aufwendungen klein zu halten, wurden einige Einschränkungen vorgenommen, welche den an eine Linking-Studie formulierten Anforderungen teilweise zuwiderlaufen:

- Als Ankeraufgaben wurden aus der Normierungsstudie lediglich Aufgaben mit einer Bearbeitungszeit von 80 Minuten in einem Pilotierungsheft verwendet. Diese können die facettenreiche Mathematik natürlich nur deutlich unvollständiger abbilden.
- Die Stichprobe der Schulen wird nicht von Experten gezogen, sondern durch die Verwaltungen in den Ländern, allerdings nach zentralen Vorgaben.
- Da die Vergleichsarbeiten in der dritten und achten Jahrgangsstufe geschrieben werden, wurde die Belastung der durchführenden Landesinstitute dadurch reduziert, dass immer nur die (identische) Hälfte der 16 Länder in die Pilotierung einbezogen wurde, was die Repräsentativität beeinträchtigt.
- Die Durchführung der Linking-Studie regelt jedes Land selbst. So werden durch das ISQ für Berlin und Brandenburg tatsächlich auch durchführende Testleitungen rekrutiert und geschult. Das Vorgehen des ISQ sichert damit ab, dass die Aufgaben der

ergänzen die Operationalisierung des Bereichs Mathematik. Für die Folgeerhebung 2018 wurden teilweise neue Aufgaben entwickelt, so dass sich die Bearbeitungszeit sogar auf 820 Minuten erweiterte.

Normierung, die hier als Anker Verwendung finden, unter etwa gleichen Bedingungen eingesetzt werden, wie in der Normierung und sich so ihre Schwierigkeit im besten Fall rekonstruieren lässt. In anderen Ländern werden die Tests allerdings von Lehrkräften selbst unter Zuhilfenahme einer Anleitung durchgeführt. Damit werden die parallel verabreichten neuen Aufgaben im gleichen Modus getestet wie in der späteren VERA-Durchführung und die im Linking bestimmten Schwierigkeitsparameter gelten eher für einen Einsatz in diesem Modus.[11]

- Die neuen Aufgaben haben vor dem Linking lediglich eine Prä-Pilotierung durchlaufen. Aus den im Rahmen dieser Linking-Untersuchung eingesetzten Aufgabenblöcken werden jene ausgewählt, deren Funktion sich bewährt. So erfüllt die eigentliche Linking-Prozedur für die neuen VERA-Aufgaben gleichermaßen die Aufgaben von Pilotierung und Normierung. Es muss also davon ausgegangen werden, dass nur zufällig ein Testheft aus dieser *Pilo-Normierung* in genau dieser Form (Itemposition und -kontext), für die die ermittelten Schwierigkeiten gelten, auch bei den Vergleichsarbeiten eingesetzt wird.

Die Abbildung 3.4 stellt die Verhältnisse der Zahl ausgewählter Aufgaben der einzelnen Untersuchungen in der horizontalen Ausdehnung dar und die Verhältnisse der Zahl ausgewählter Personen in der vertikalen. Hierbei wird deutlich, dass die Anbindung der Metrik der in den Vergleichsarbeiten eingesetzten Tests an die Metrik der Bildungsstandards nur über eine „schmale“ Linking-Erhebung realsiert wird. Dabei meint „schmal“ sowohl die relativ kleine Stichprobe von Personen und von Aufgaben. Ein Vergleich von zwei VERA-Ergebnissen stützt sich damit auf zwei solche nacheinander ablaufenden Linking-Prozesse. Die Anbindung der Erhebungen zum Bildungstrend ist demgegenüber deutlich weniger prekär.

[11] Hier schließt sich eine Diskussion an, welchem Vorgehen der Vorzug zu geben wäre. Eine Untersuchungen zu den verschiedenen Wirkungen ist nicht bekannt.

4. Überprüfung von Gewissheiten beim Einsatz der Rasch-Skalierung

In dieser ersten Studie wird untersucht, inwiefern die konkrete Interpretation von Fähigkeitsparametern aus Rasch-Skalierungen durch die dahinterliegenden Modellannahmen gestützt werden. Aus den der Konstruktion der Rasch-Skala zugrundeliegenden Modellannahmen (Abschnitt 3.3) sowie der damit verbundenen Argumentation bei der Ableitung von Kompetenzstufen (Abschnitt 3.4) ergeben sich vier Schlussfolgerungen, die im Folgenden als Hypothesen formuliert und überprüft werden. Diese Nachweise sind notwendig zu stützende Elemente der *Skalierung* im Validitätsargument (vergleiche mit Abbildung 2.1), also der Schlussfolgerung vom beobachteten Antwortvektor auf das beobachtete Ergebnis.

4.1. Gewissheiten

Dass Modelle die Realität nur hinreichend gut beschreiben ist unbestritten und dass das bei aller mathematischen Komplexität doch eher triviale Rasch-Modell hier keine Ausnahme bildet, wurde schon ausgeführt. Im Rahmen dieses probabilistischen Modells sind einige Schlussfolgerung von besonderer Relevanz, rahmen geradezu die Anwendung der Modelle in der Praxis und bilden in diesem Sinne Gewissheiten der an der Testentwicklung Beteiligten ab. Noch grundlegender ist, dass diese hier als Hypothesen formulierten Gewissheiten durch die Kommunikation im Rahmen der Vergleichsarbeiten ein Grundgerüst bilden, vor dem Ergebnisse letztendlich von der Schulpraxis interpretiert werden.

Die folgenden Abschnitte leiten vier Hypothesen zur Validität von Fähigkeitsschätzungen aus der Konstruktion der Rasch-Skalierung ab und beziehen sich dabei jeweils auf praktische Kontexte. Am Ende jedes Unterpunktes wird die Hypothese zur Prüfung ausformuliert. Nach der Beschreibung der Methodik und der zum Nachweis der Hypothesen verwendeten Daten, erfolgt die Prüfung der Hypothesen und abschließend eine Zusammenfassung der Ergebnisse.

4.1.1. Irrelevanz der Itemauswahl

Die mittleren Schwierigkeiten der Testhefte des IQB changieren für eine Version über die Jahre wie auch für die Abstände zwischen den Versionen innerhalb eines Durchgangs. Die Schwierigkeitsabstufungen zwischen den Testheftversionen eines Durchgangs sollen dabei eine Passung zu den erwarteten Leistungen der unterschiedlichen Schulformen und Bildungsgänge herstellen (zum Beispiel Aneis et al., 2018, S.14).

Bei ungünstiger Passung von Itemschwierigkeiten und Personenfähigkeiten, wenn also die mittlere Lösungshäufigkeit der Items eines Testhefts für eine Teilpopulation substantiell von 50% abweicht, werden die Standardfehler der Schätzungen größer.

$$s(E_\theta) = \frac{1}{\sqrt{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})}} \tag{4.1}$$

Ist die Differenz zwischen Itemschwierigkeit und Personenfähigkeit gleich Null, wird einer Person also ein Item vorgelegt, dessen Schwierigkeit exakt der Fähigkeit der Person entspricht, ergibt sich nach der Festlegung (vgl. Abschnitt 3.3) eine Lösungswahrscheinlichkeit von 50% also $p_{vi} = 0{,}5$. Hier ist $p_{vi}(1 - p_{vi})$ maximal und für größere Differenzen von Itemschwierigkeit und Personenfähigkeit stets kleiner. Für einen kleinen Standardschätzfehler sollte die Itemschwierigkeit also dort liegen, wo die Personenfähigkeit erwartet wird. Verfolgt man ausschließlich das Ziel einer optimalen Fähigkeitsschätzung, wäre ein für die jeweilige Fähigkeit der Person angepasstes Testheft optimal. Ideal, ein Verabreichen von Items mit genau der Schwierigkeit, die der Fähigkeitsschätzung auf der Basis der zuvor verabreichten Items entspricht (adaptives Testen). Sofern man nicht vollständig mit automatisch bewertbaren Aufgaben online testen kann, ist eine Umsetzung kaum denkbar. Zudem besteht im Zusammenhang mit den Vergleichsarbeiten nicht nur das ausschließliche Ziel der Schätzung des einen Fähigkeitsparameters. Rückmeldungen berichten vielschichtig über Lösungen von Einzelaufgaben oder Aufgabengruppen. Wird einer Person ein Testheft unangemessener Schwierigkeit vorgelegt, vergrößert sich lediglich der Schätzfehler. Im Mittel sollten die Schätzungen aber weiterhin korrekt sein.

Bei einem aus psychometrischer Sicht optimalen Testheft für eine Gruppe von Schüler*innen mit einer bestimmten Fähigkeitsverteilung, sollte der Mittelwert der Itemschwierigkeiten dem Mittelwert der Personenfähigkeiten entsprechen. Das Testheft weist dann in der untersuchten Population eine Lösungshäufigkeit von 50% auf. Damit begründet sich die Zuordnung von verfügbaren Testheftalternativen auf konkrete Schulformen genauso wie eine Auswahl

angepasster Testhefte auf Ebene der Klasse oder der Schüler*innen.

Wie schon aus der Theorie der Rasch-Skalierung als Vorteil herausgestellt, führt die Zusammenstellung eines Testhefts mit unterschiedlichen, insbesondere auch unterschiedlich schwierigen Items immer zu (auf jeden Fall im Mittel) identischen Schätzungen der Personenparameter.

Nicht auszuschließen ist, dass Schüler*innen beispielsweise auf Grund von Demotivation bei zu schwierigen Tests, von Unterforderung bei zu einfachen Tests oder bei einer ungünstigen Anordnung, identische Items mit weniger Erfolg bearbeiten, als sie dies bei günstigerer Konstellation geschafft hätten. Solche Effekte, wenn auch praktisch relevant, sollen im Folgenden keine Rolle spielen. Aus der theoretischen Konstruktion der Rasch-Skalierung ergibt sich die

Hypothese 1: Unabhängig von der Auswahl der Items für ein Testheft, ergeben sich für eine Person nur zufällig verschiedene, im Mittel identische Fähigkeitsschätzer.

4.1.2. Erwartungstreue Schätzung von Personenparametern

Für den im Rahmen der Vergleichsarbeiten vorliegenden Fall bekannter Itemparameter findet zur Schätzung der Personenparameter üblich die WLE-Methode (Weighted-Likelihood-Estimates) nach Warm (1989) Anwendung. Diese Schätzung ist erwartungstreu. Sie führt auch für Probanden, die keine bzw. alle Items richtig gelöst haben zu einer plausiblen Schätzung, bzw. zu einer plausiblen Extrapolation, denn für eine Person die alle Items eines Tests richtig löst, lässt sich die Fähigkeit lediglich nach unten mit hoher Wahrscheinlichkeit einschränken. Für eine Abschätzung nach oben gibt es keine Information. Dies gilt invers für den Fall von ausschließlich falsch gelösten Items (siehe auch Abschnitt 3.3). Die Extrapolation ist aber dennoch hilfreich. Im Allgemeinen ist in einem solchen Fall eine Positionierung der Personenfähigkeit in der extremen Randlage ausreichend.

Zudem zeigt sich bei der Berechnung des Personenparameters als Bestimmung des Maximums der bedingten Wahrscheinlichkeiten äquivalent zur Gleichung 3.3, dass sich für eine bestimmte Zahl an richtig gelösten Items immer ein identischer Personenparameter ergibt. Rost (2004, 122ff) leitet mathematisch ab, dass die sogenannten Randsummen sämtliche Informationen des Modells enthalten. Die Summe richtig gelöster Items wird deshalb als suffiziente, also erschöpfende Statistik bezeichnet.

Die Schätzung der Personenparameter ergibt für alle Personen mit p richtig gelösten Items den gleichen Wert, unabhängig davon, welche konkreten Items korrekt gelöst wurden. Aus schulpraktischer Sicht ist dieser Umstand mindestens erklärungsbedürftig, wo doch die unter-

schiedlichen Schwierigkeiten der Items ein zentrales Element der Kommunikation darstellen. Dass dann aber eine richtige Antwort auf ein sehr einfaches Item den gleichen Beitrag zur Fähigkeitsfeststellung leistet wie ein richtig gelöstes schwieriges Item, widerspricht erst einmal einer schulpraktischen Sicht, bei der man die richtige Lösung einer schwierigen Aufgabe eher mit mehr Punkten belohnt. Die Verwertung dieses Ergebnisses der Vergleichsarbeiten ist aber originäre Aufgabe der Schulpraxis.

Es ergeben sich damit bei einer Skalierung nach Rasch für ein Instrument mit k Items rechnerisch lediglich $k+1$ mögliche Personenparameter. Durch die Messung wird die Fähigkeit also nur an $k+1$ Stellen auf der an sich rationalen Skala der Personenfähigkeiten abgebildet[1]. Dies führt zur Annahme: Der durch die Rasch-Skalierung zugeordnete Personenparameter liegt in direkter Nähe des wahren Personenparameters. Es gibt keinen zweiten zuordenbaren Personenparameter, der dichter liegt.

Der Standardfall ist sicherlich eine Antwortmatrix, die deutlich mehr Personen (N) enthält als Items (k), also $N \gg k$ gilt. Die Folge ist, dass für die Schätzung eines Itemparameters deutlich mehr Messwerte zur Verfügung stehen, als für die Schätzung eines Personenparameters. Bei den Vergleichsarbeiten liegt der Fokus allerdings genau auf der Schätzung der Personenparameter bei bekannten, festen Itemparametern. Wegen der deshalb deutlich größeren Messfehler für Personenparameter wird die Hypothese etwas konservativer formuliert:

Hypothese 2: Der Mittelwert der wahren Personenparameter aller durch die Rasch-Skalierung einem gemessenem Personenparameter zugeordneten Personen liegt in direkter Nähe dieses Personenparameters. Es gibt keinen zweiten möglichen Personenparameter, der dichter liegt.

4.1.3. Die Bedeutung von Guttman-Pattern

Wird die Antwort einer Schülerin auf das i-te Item mit x_i beschrieben, nennt man die gemeinsame Darstellung aller k Items für diese Schülerin einen Vektor $\bar{x}$ der Länge k. Wenn dieser Vektor alle Antworten für eine Schülerin oder einen Schüler umfasst, wird dieser Vektor auch Antwortpattern genannt. Es enthält für jedes Item lediglich das Ergebnis, kodiert als richtig (1) vs. falsch (0).

Ein Antwortpattern, welches bei einer nach Schwierigkeit geordneten Itemmenge von k Items mit p ($p > 0$ und $p < k$) korrekt bearbeiteten Items beginnt und ab einem Punkt $p+1$ nur noch $(k-p)$ falsch bearbeitete Items aufweist, wird folgend Guttman-Pattern S_p^k

[1] Das Instrument diskretisiert die Skala für die Messung so, wie ein Lineal die an sich rationale Skala der Streckenmessung nur an endlich vielen Punkten abbildet.

genannt[2]. Für jedes p existiert genau ein Guttman-Pattern. Mit zunehmender Itemzahl k steigt die Anzahl möglicher Pattern mit p korrekt gelösten Items auf

$$N_{pattern} = \frac{k!}{(k-p)!} \tag{4.2}$$

Damit sinkt schon bei völlig zufälligem Auftreten die Wahrscheinlichkeit jedes einzelnen und damit auch des einen Guttman-Pattern (siehe Tabelle 4.1, Spalten „Anzahl mögl. Pattern“ sowie „mögl. Pattern^{-1}“). Tatsächlich sind die verschiedenen Pattern aber nicht gleichwahrscheinlich. Die Wahrscheinlichkeit für die korrekte Lösung eines Items i ergibt sich nach der Formel 3.1 zu

$$P(x_i = 1) = \frac{e^{(\theta-\sigma_i)}}{1 + e^{(\theta-\sigma_i)}} \tag{4.3}$$

und für eine falsche Lösung

$$P(x_i = 0) = \frac{1}{1 + e^{(\theta-\sigma_i)}} \tag{4.4}$$

Die Modellgleichung für das i-te von k Items kann als

$$P(x_i = 1) = \frac{e^{x_i(\theta-\sigma_i)}}{1 + e^{(\theta-\sigma_i)}} \tag{4.5}$$

zusammengefasst werden. Der Vektor $\bar{x}$ ist hier das Antwortpattern einer Person für die k Items $\bar{x} = (x_1, x_2, ...x_i...x_k)$. Für eine Person mit der Fähigkeit θ ergibt sich die bedingte Wahrscheinlichkeit für ein bestimmtes Antwortpattern $\bar{x}$ wegen der vorausgesetzten Unabhängigkeit aller Items als Produkt der Wahrscheinlichkeiten für jedes Item, also zu

$$P(\bar{x}) = \prod_{i=1}^{k} \frac{e^{x_i(\theta-\sigma_i)}}{1 + e^{(\theta-\sigma_i)}} \tag{4.6}$$

Welches Pattern hat für eine Person die größte Wahrscheinlichkeit? Für das Produkt von Brüchen kann man auch schreiben

$$P(\bar{x}) = \frac{\prod_{i=1}^{k} e^{x_i(\theta-\sigma_i)}}{\prod_{i=1}^{k} 1 + e^{(\theta-\sigma_i)}} \tag{4.7}$$

Der Nenner ist unabhängig vom Antwortpattern, also konstant. Für nicht gelöste Items ist der Zähler mit $x_i = 0$ und damit $e^0 = 1$. Für richtig gelöste Items ist die Differenz von Personenfähigkeit θ und Itemschwierigkeit σ_i entscheidend. Hierbei gilt: Ist die Schwierigkeit

[2] Für einen beispielhaften Test mit 48 Aufgaben, von denen 13 korrekt gelösten wurden ist das Guttman-Pattern jenes, bei dem genau die 13 leichtesten Aufgaben korrekt gelöst wurden und die 35 schwersten Aufgaben gerade nicht.

Tabelle 4.1.: Häufigkeiten und Auftretenswahrscheinlichkeiten von Antwortpattern für ein Testheft mit 48 Items

korrekt gelöste Items (p)	Anzahl möglicher Pattern	P eines Pattern^{-1}	Wahrscheinlichkeit für ein Guttman-Pattern $P\|S_p^k$
0	1	1	-
1	48	0.0208	0.1174
2	1128	8.9×10^{-4}	1.9×10^{-2}
...			
13	1.9×10^{11}	5.2×10^{-12}	2.1×10^{-6}
...			
23	3.1×10^{13}	3.2×10^{-14}	5.2×10^{-7}
24	3.2×10^{13}	3.1×10^{-14}	4.6×10^{-7}

σ_i kleiner als die Personenfähigkeit θ, wird der Faktor wegen $\theta - \sigma_i > 0$ im Exponenten der e-Funktion größer 1 und im umgekehrten Fall kleiner 1. Die Wahrscheinlichkeit für ein Antwortpattern ist also dann am größten, wenn Items mit Schwierigkeiten kleiner als die Personenfähigkeit richtig gelöst werden und solche, deren Schwierigkeit größer als die der Personenfähigkeit ist ungelöst bleiben.

Dies genau beschreibt das Guttman-Pattern S_p^k und zudem die triviale wie plausible Erwartung, dass die geschätzte Personenfähigkeit einer Person, die ein solches Pattern als Antwortverhalten aufzeigt, zwischen dem schwierigsten richtig gelösten und dem einfachsten nicht richtig gelösten Item liegt. Von allen Pattern mit p korrekt gelösten Items, ist das eine Guttman-Pattern S_p^k jenes, mit der größten Wahrscheinlichkeit. In Tabelle 4.1 wurden einige Wahrscheinlichkeiten für ein Beispiel mit 48 Items berechnet (VERA-8 Mathematik, 2015, Testheft A). Die erste Zeile mit $p = 0$ ist nur illustrierend, denn für ein Guttman-Pattern müssen p wie k größer 0 sein.

Wegen der geringen Wahrscheinlichkeit verwundert das tatsächlich seltene Vorkommen von Guttman-Pattern selbst in großen Stichproben von Large-Scale-Assessments (LSA) mit einigen Zehntausend Schüler*innen ebenso wenig, wie ihr vollkommenes Ausbleiben. Beispielhaft für die Itemverteilung des Tests mit $k = 48$ Aufgaben und mit genau $p = 13$ korrekt gelösten Aufgaben, ergeben sich ca. 1.9×10^{11} mögliche Antwortpattern und damit eine unbedingte Wahrscheinlichkeit von ca. 5.2×10^{-12}. Das ist die Wahrscheinlichkeit für ein bestimmtes Pattern mit 13 von 48 korrekt gelösten Aufgaben, wenn alle möglichen dieser Pattern gleich wahrscheinlich wären. Tatsächlich sind die Wahrscheinlichkeiten aber für unterschiedliche Pattern verschieden und die Wahrscheinlichkeit für das beispielhafte Guttman-Pattern mit 2.1×10^{-6} klein, aber immerhin gut 400.000 Mal größer als die unbedingte Wahrscheinlich-

Tabelle 4.2.: Erwartete und reale Häufigkeit von Guttman-Pattern bei VERA-8 Mathematik (Berlin, 2015)

				Guttman-Pattern	
richtig gelöst	abs.	rel.	erwartet rel.	erwartet abs.	realisiert abs.
0	14	0.13%	-	-	-
1	6	0.05%	11.47%	0.69	1
2	8	0.07%	0.02%	0.00	0
...					
13	142	1.28%	0.00%	0.00	0
...					
41	289	2.61%	0.03%	0.10	0
42	218	1.97%	0.10%	0.22	0
43	172	1.55%	0.34%	0.58	0
44	134	1.21%	1.14%	1.53	2
45	91	0.82%	2.25%	2.05	1
46	59	0.53%	6.17%	3.64	5
47	26	0.23%	17.70%	4.60	6
48	15	0.14%	-	-	-
Summe	11079			13.41	15

keit. Das Testheft mit den 48 Items wurde in Berlin 11.079 Mal eingesetzt. 13 von 48 korrekt gelösten Items fanden sich dabei genau 142 Mal, ein Guttman-Pattern fand sich dabei nicht. (siehe Tabelle 4.2). Die Tabelle zeigt, dass Guttman-Pattern nur an den Rändern eine erwartete Häufigkeit aufweisen, die eine Realisierung mit einer angemessenen Wahrscheinlichkeit erwarten lässt. Entsprechend der erwarteten Häufigkeiten im Beispiel sollten gut 13 Guttman-Pattern zu finden sein, tatsächlich sind es 15.

Ungeachtet dessen weist die höchste Auftretenswahrscheinlichkeit dem Guttman-Pattern eine besondere Bedeutung zu. Denn die Vorstellung vom Guttman-Pattern nimmt beim Standard-Setting eine zentrale Rolle ein. Den Expert*innen wird beim Standard-Setting zum Auffinden der Kompetenzstufengrenzen mit Hilfe des OIB (siehe 3.4) eine Anweisung gegeben, die für das Beispiel der Stufengrenze zwischen A1 und A2 im GER für die Fremdsprache Englisch von Harsch et al. (2010, S. 92) wie folgt wörtlich zitiert und äquivalent auch bei Tiffin-Richards et al. (2013, S. 18) beschrieben wird:

> „We are looking for ‚mastery'. Mastery is defined as the point where the minimally proficient student for a given proficiency level would have better than a two-thirds likelihood of answering an item correctly (response probability = .67). Think of your least proficient A2 student. Start at Item 1; ask yorself if the student you have in mind has better than two-thirds of a chance of getting that answer correct.

> If the answer is ‚yes', move to the next most difficult item. Follow this procedure until you reach the item where you think this student has less than a two-thirds chance of getting the item correct. Place the A2 bookmark ON TOP of the page where that item is located."

Dass als Antwortwahrscheinlichkeit (response probability) 2/3 verwendet wird und nicht wie im Rasch-Modell 1/2, sorgt lediglich für eine bessere Einschätzung der Expert*innen und wird letztendlich wieder mit

$$Logit^* = Logit + ln(\frac{P_{rp}}{1 - P_{rp}}) \tag{4.8}$$

auf eine Wahrscheinlichkeit von 0,5 zurückgerechnet, wobei *Logit* den ursprünglichen Wert mit Bezug zu 0,5 meint und $Logit^*$ den für die Antwortwahrscheinlichkeit von P_{rp} transformierten Wert.

Pant et al. (2010) stellen fest, dass Validitätsbetrachtungen sowohl das Gesamtsystem des Large Scale Assessments wie auch dessen Subsysteme betreffen sollten und halten fest, dass die Festlegung der Cut-Scores „ein besonders kritisches Verbindungsglied zwischen dem evidenzbezogenen, empirisch gut untersuchbaren Aspekten des Gesamtsystems und den konsequenzbezogenen, eher normativen und praxisrelevanten Aspekten" (ebenda, S.178) darstellt. Wenn der Vorstellung des Guttman-Pattern an dieser kritischen Stelle eine so essentielle Bedeutung zugeschrieben wird, wird sich diese in der Interpretation niederschlagen.

Für die dritte Hypothese wird hier definiert:

> *Definition*: Eine Metrik ist genau dann gegenüber Guttman-Pattern unverzerrt, wenn alle Personenfähigkeiten von Personen, deren Antworten ein Guttman-Pattern darstellen zwischen dem schwierigsten richtig gelösten und dem leichtesten nicht richtig gelöstem Item liegen.

Zieht man in Betracht, dass sich Guttman-Pattern auch zufällig ergeben können, kann man diese Definition in folgender Form abschwächen:

> *Definition*: Eine Metrik ist genau dann gegenüber Guttman-Pattern unverzerrt, wenn der Mittelwert aller Personenfähigkeiten von Personen, deren Antworten ein konkretes Guttman-Pattern darstellen zwischen dem schwierigsten richtig gelösten und dem leichtesten nicht richtig gelöstem Item liegen.

Auch, wenn man für das probabilistische Rasch-Modell akzeptiert, dass das Vorkommen eines Guttman-Pattern selten ist, kann man mit Argumenten sowohl aus der psychometrischen

Konstruktion der Rasch-Skalierung wie auch aus der empirischen Validierung des Kompetenzstufenmodells die dritte Hypothese begründen:

Hypothese 3: Eine mit Hilfe der Rasch-Skalierung definierte Metrik, ist gegenüber Guttman-Pattern unverzerrt.

4.1.4. Zusammenhang der Verteilung von Itemschwierigkeiten und Personenfähigkeiten

Aus der Informationsfunktion für einen Test

$$I(\delta) = \sum_{i=1}^{k} p_{vi}(1 - p_{vi}) \tag{4.9}$$

kann man entnehmen, dass ein Rasch-Skalierter Test nicht an jeder Stelle mit gleicher Genauigkeit misst. Da die Informationsfunktion das Reziprok der Wurzel des Standardfehlers ist

$$I(\delta) = \frac{1}{\sqrt{s(E_\delta)}} = \frac{1}{Var(E_\delta)} \tag{4.10}$$

kann man dafür an die Argumentation bei Formel 4.1 anschließen. Dort wurde festgestellt, dass für einen minimalen Standardfehler $s(E_\delta)$ und damit für die maximale Information $I(\delta)$ der Mittelwert der Itemschwierigkeiten dem Mittelwert der Personenparameter entsprechen sollte. Tatsächlich ist der Standardfehler für Schüler*innen kleiner, wenn deren Fähigkeit nah am Mittelwert der Itemschwierigkeiten und groß, wenn die Fähigkeit weiter am Rand liegt. Formel 4.9 verdeutlicht aber auch, dass jedes Item einen Beitrag zur Ermittlung jedes Personenparameters leistet. Mit größerem Abstand der Items von der Position des Personenparameters wird dieser Beitrag, den die Informationsfunktion aufsummiert, kleiner. Folgerichtig entscheiden also die Positionen der Items auf der Skala der Schwierigkeiten darüber, in welchem Intervall besonders genau gemessen wird.

In der Praxis findet die Anpassung von Itemschwierigkeiten zur Fähigkeit der untersuchten Population in zwei Schritten statt. Auf der Basis der Ermittlung der empirischen Schwierigkeit aller Items im Rahmen einer Pilotierung konstruiert das IQB Testhefte durch eine Itemauswahl „unter gleichzeitiger Berücksichtigung didaktischer und psychometrischer Gesichtspunkte“ (Penk et al., 2014). Für VERA-8 wurden bis 2015 drei Testhefte zur Verfügung gestellt. Das einfachste Testheft A fokussierte auf Schüler*innen, die erwartungsgemäß einen Hauptschulabschluss anstreben. Für Schüler*innen, die den Mittleren Schulabschluss anstreben wurde das Testheft B konzipiert. Das schwierige Testheft C zielte auf Schüler*innen,

die ein Abitur erreichen wollen. Nach 2015 wurden für VERA-8 nur noch zwei Testhefte für Gymnasien und andere Schulformen ausgeliefert. Für VERA-3 wurde ein Testheft an die Population insgesamt angepasst. Alle Testhefte sollten dabei bundesweit in ihrer jeweiligen Population mittlere Lösungshäufigkeiten zwischen 50 und 60% aufweisen[3]. Dazu wurden für jedes Testheft Items ausgewählt, die sich entsprechend auf die Kompetenzstufen verteilen. Am deutlichsten aber wird die Idee der Anpassung der Messgenauigkeit des Instruments an eine Zielpopulation im Rahmen der 2020 einsetzenden Modularisierung (Aneis et al., 2020, S.14-16). Durch eine Häufung von Items bestimmter Schwierigkeit in spezifischen Intervallen der Fähigkeitsskala sollen Testmodule bestimmte Schwierigkeitsverteilungen aufweisen. Im Fall des Fachs Mathematik bei VERA-8 passiert dies sowohl beim Basis- wie beim Ergänzungsmodul. Ein Basismodul soll obligatorisch allen Schüler*innen vorgelegt werden. Hier gab es 2020 das erste Mal zwei Module mit unterschiedlicher Schwierigkeitsverteilung: in beiden Modulen verteilen sich die Aufgaben über alle fünf Leitideen, das Basismodul A soll dabei aber eher den unteren bis mittleren Kompetenzbereich abdecken, während das Basismodul B eher auf den mittleren und oberen Bereich fokussiert. Für die Ergänzungsmodule wird konkreter formuliert, dass das einfachere Modul A wieder im unteren und mittleren Bereich stärker differenzieren soll und dafür primär Aufgaben der Kompetenzstufen I, II sowie III und vereinzelt solche der Kompetenzstufe IV und V enthalten soll. Das schwierigere Ergänzungsmodul B enthält primär Aufgaben der Kompetenzbereiche III bis V und vereinzelt aus dem Bereich II und I und differenziert damit eher im mittleren bis oberen Kompetenzbereich. Festzuhalten ist: Durch eine Auswahl von Items entsprechend ihrer empirischen Schwierigkeit, soll die Messgenauigkeit in einem bestimmten Fähigkeitsbereich erhöht werden.

In einem zweiten Schritt muss nun die Administration jedes Landes über den Einsatz dieser Testhefte entscheiden. Vor der Modularisierung blieb überhaupt nur deshalb Spielraum, weil das Schulsystem Berlins, wie auch in einigen anderen Ländern, seit 2010 nur noch zweigliedrig ist, aber bis 2015 Testhefte für drei Schulformen bereitgestellt wurden. In erster Linie wurde versucht für jede Schulform ein Testheft so auszuwählen, dass die schulspezifische Lösungshäufigkeit in etwa zwischen 50 und 60 Prozent lag, auf jeden Fall nicht darunter. Durch die Modularisierung ab 2020 müssen neue Verfahren für die Auswahl der Module verschiedener Schwierigkeiten entwickelt werden, insbesondere, weil die Auswahl vermutlich nicht mehr zentral, sondern durch die einzelne Schule ggf. die einzelne Lehrkraft getroffen wird.

[3]Diese Anforderung wurde so von den Ländern formuliert und im öffentlich nicht verfügbaren Lasten- und Pflichtenheft festgehalten. Im technischen Report, beispielsweise bei Aneis et al. (2018, S.16), werden die mittleren Lösungshäufigkeit angegeben, die sich durch die Aufgabenkombinationen für die einzelnen Testheftvarianten ergaben.

Hierbei ist die itembezogene Konstruktion aus dem ersten Schritt selbstverständlich entscheidungsrelevant. Grundlage für dieses Vorgehen ist offenbar die in Hypothese 4 beschriebene Gewissheit.

Hypothese 4: Durch eine fokussierte Auswahl von Items in einem Schwierigkeitsintervall kann die Verteilung des Messfehlers eines Instruments zielgerichtet beeinflusst werden, so dass es in diesem Intervall besonders sicher misst, folglich dort besonders gut differenziert.

4.2. Methode

Ergebnisse von Vergleichsarbeiten stehen nicht selten in der Kritik. So wurde vermutet und auch untersucht, dass zum Beispiel der frühzeitige Versand der Testhefte an die Schulen sowie die eigenständige Durchführung und Kodierung durch die Lehrkräfte gegenüber externen Testleitungen die gemessenen Leistungen der Schülerinnen und Schüler positiv verzerrt (zum Beispiel Graf et al., 2013). Solche Aspekte sollen für die vorliegende Untersuchung ausgeschlossen werden. Deshalb wird zur Prüfung der Hypothesen zu den vier Gewissheiten das Antwortverhalten von Personen simuliert, denen sämtliche bei VERA-8 genutzten Testhefte „vorgelegt“ werden. Alle Berechnungen wurden für die Testdomänen *Mathematik*, *Deutsch Lesen* sowie *Englisch Lesen* durchgeführt. Die folgenden Beschreibungen beziehen sich immer auf die Ausführungen für die Domäne Mathematik. Informationen zu ggf. abweichenden Ergebnissen für die anderen Domänen werden im Text erwähnt, Ergebnisdaten im Anhang dargestellt.

Im Standardfall (siehe die übliche Testanordnung im oberen Teil der Abbildung 4.1) bearbeiten Personen mit ihren wahren aber unbekannten Personenparametern ein Testheft, das aus einem Set an Items (Testheftversion) besteht. Die Antwortmatrix wird mit den bekannten Parametern der Itemschwierigkeit der Rasch-Skalierung übergeben, deren Ergebnis die Schätzung der Fähigkeitspartameter ist. In der Simulation (Abbildung 4.1, unterer Teil) werden die wahren unbekannten Personenparameter durch simulierte bekannte Parameter ersetzt und die Bearbeitung der Items durch diese simulierten Personen wird auf der Basis der Itemparameter und der Annahmen des Rasch-Modells in einem als *inverse Rasch* bezeichneten Algorithmus simuliert. Wird beispielsweise 20 Personen, denen durch die Simulation einen Fähigkeitsparameter von 430 zugeordnet wurde, ein Item mit einer Schwierigkeit von 500 vorgelegt, dann ergibt sich aus der ICC für die Lösungswahrscheinlichkeit ein Wert von beispielsweise 35%. Für zufällig ausgewählte 35% der 20 Personen, also für 7, wird das Item auf richtig gelöst gesetzt und für die anderen 13 auf falsch. Aus dieser Simulation der Testheftbearbeitung folgt

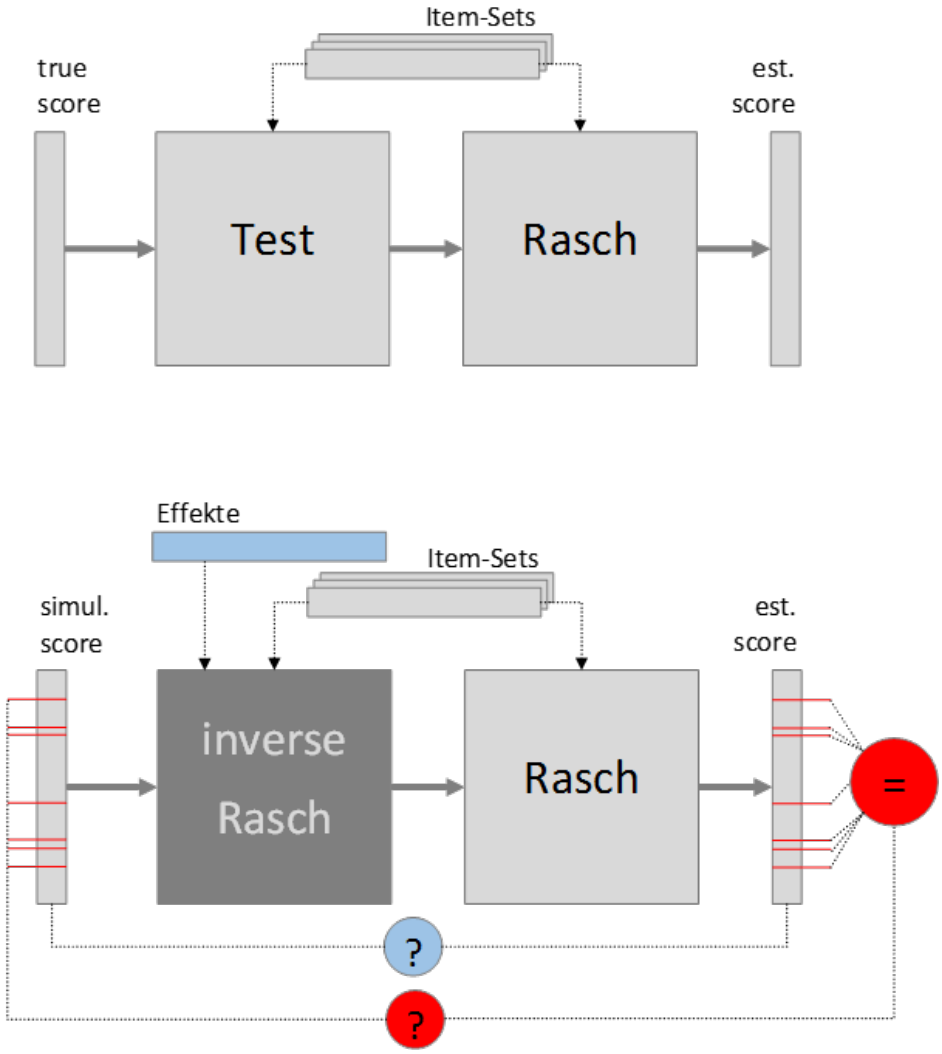


Abbildung 4.1.: Übliches Messmodell (oben) und Variation zur Hypothesenprüfung (unten)

eine Rasch-konforme Antwortmatrix.

Die Schätzung der Personenparameter erfolgt dann wie zuvor über eine Rasch-Skalierung. Weil nun die wahren Parameter bekannt sind, kann die Differenz zwischen der Schätzung und den simulierten wahren Werten als Fehler bestimmt werden. Ggf. kann das *Antwortverhalten* der simulierten Personen durch gezielte Effekte beeinflusst werden. So kann abgeschätzt werden, wie sich diese Effekte auf die Güte der Schätzung auswirken (blau). Konkret wird weiterhin betrachtet, wie sich solche simulierten wahren Werte verteilen, deren Schätzwert identisch ist (rot).

4.3. Daten

Wie aus Abbildung 4.1 abzuleiten ist werden zweierlei Daten benötigt: die Schwierigkeiten von Items aus Testheften (Item-Sets) sowie simulierte Personenfähigkeiten. In den folgenden zwei Abschnitten werden diese Daten beschrieben.

4.3.1. Itemparameter aus den VERA-Durchgängen

Für VERA-8 liegen für die Jahre 2008 bis 2015 sämtliche Itemparameter für jeweils drei und für die Folgejahre bis 2020 für jeweils zwei vom IQB zur Verfügung gestellte Testheftversionen vor. Die Tabelle 4.3 gibt einen Überblick über die Verteilung der Itemparameter für jedes der 34 Testhefte. Während alle anderen Werte aus den Itemkennwerttabellen des IQB zu entnehmen waren, lagen für die durchschnittliche Diskrimination und ihrer Standardabweichung nur Werte aus der eigenen VERA-Durchführung vor. Wegen des erwartbar deutlich größeren Fehlers blieben die Einsätze von Testheften in kleinen Teilstichproben unberücksichtigt, wie der Einsatz der Testheftversionen 3 in Mathematik-Profilklassen oder für Englisch in bilingual unterrichteten Klassen.

Tabelle 4.3.: Beschreibung der VERA-8-Testinstrumente für das Fach Mathematik

Ver.[a]	Jahr	N	Itemparameter						Testhefte				
			BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung				
			Min	Max	Mw	Sd	Mw	Sd	I	II	III	IV	V
1	2008	57	112	699	404	139	-	-	28	14	8	4	3
	2009	47	143	775	456	118	-	-	13	15	12	6	1
	2010	33	167	684	461	126	1.26	0.31	10	8	9	5	1
	2011	39	119	679	349	141	1.15	0.36	23	7	5	3	1
	2012	38	124	603	355	132	1.12	0.28	26	5	4	3	0
	2013	42	60	512	334	96	-	-	32	9	1	0	0
	2014	43	67	656	365	128	1.20	0.33	29	10	2	2	0
	2015	48	159	655	378	119	1.13	0.27	33	10	1	4	0
2	2008	56	201	752	439	154	1,04	0,33	25	13	6	4	8
	2009	48	143	794	497	145	1.34	0.39	12	11	13	6	6
	2010	36	210	976	506	154	1.48	0.50	9	10	5	7	5

Fortsetzung auf der nächste Seite

Tabelle 4.3.: Fortsetzung der Tabelle von der Vorseite

Ver.[a]	Jahr	N	Itemparameter						Testhefte				
			BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung				
			Min	Max	Mw	Sd	Mw	Sd	I	II	III	IV	V
	2011	38	122	681	459	155	0.93	0.35	15	7	4	9	3
	2012	36	174	716	436	139	0.95	0.33	16	6	7	6	1
	2013	42	60	722	421	141	1.13	0.39	21	11	3	4	3
	2014	38	235	765	455	133	1.10	0.43	17	10	6	2	3
	2015	44	229	728	463	149	0.93	0.35	21	8	4	5	6
	2016	46	182	666	400	115	1.15	0.47	27	11	7	1	0
	2017	41	161	681	400	126	1.32	0.34	26	6	7	1	1
	2018	45	172	692	405	113	1.28	0.36	27	12	4	1	1
	2019	48	184	641	415	115	1.16	0.36	29	9	7	3	0
	2020	35	188	715	423	135	1.23	0.44	23	3	5	2	2
3	2008	42	201	791	528	156	-	-	8	9	8	5	12
	2009	40	143	794	560	150	-	-	4	7	12	7	10
	2010	36	188	976	533	173	-	-	8	8	6	7	7
	2011	32	281	774	546	119	-	-	4	5	8	10	5
	2012	34	174	941	553	171	-	-	6	7	5	9	7
	2013	43	235	886	525	144	0.97	0.31	9	13	5	7	9
	2014	38	341	765	550	112	-	-	9	6	8	9	6
	2015	38	323	766	557	121	-	-	8	5	10	7	8
	2016	43	390	823	556	107	0.93	0.28	8	8	12	9	6
	2017	38	316	797	562	120	1.11	0.33	6	8	10	7	7
	2018	47	305	778	554	116	1.08	0.36	8	9	10	13	7
	2019	52	318	776	559	104	1.03	0.30	7	12	13	14	6
	2020	29	332	826	550	122	1.15	0.30	7	5	6	6	5

[a] Version des Testhefts. Ab dem Jahr 2020 wurden vom IQB nicht genau zwei Testheftversionen zur Verfügung gestellt, sondern kombinierbare Module. Hier wird jeweils die empfohlene Auswahl des Landes Berlin für die ISS (Testheftversion 2) und die Gymnasien (Testheftversion 3) dargestellt.

[b] Die Diskrimination wurde aus den Daten des regulären Einsatzes bei VERA ermittelt und liegt deshalb für nicht oder nur an kleinen Teilpopulationen eingesetzten Testheftversionen nicht vor. Sie gilt darüber hinaus für die Länder Berlin und Brandenburg und kann für andere Länder entsprechend abweichen.

In dieser und folgenden Darstellungen wurden die nur noch zwei Testheftversionen ab 2016 Testheft 2 und 3 zugeordnet. Dies bildet eher die Intention der Entwickler*innen ab, nach der das jeweils schwerste Testheft nach wie vor auf Schüler*innen abzielt, die einen gymnasialen Abschluss anstreben. Dieses Testheft ist nun durchgehend Version 3.

Im Land Berlin wurden im Rahmen von VERA-8 in jeder getesteten Domäne mindestens zwei Testhefte jährlich verpflichtend eingesetzt. Somit liegen für die Mehrzahl der Items Daten aus den Vollerhebungen an Berliner Schüler*innen vor. Wegen der über die Jahre stabilen Verbindlichkeit des Einsatzes im Land Berlin konnten die folgenden Analysen für

- das Fach Mathematik,
- die Domäne Englisch Leseverstehen sowie
- die Domäne Deutsch Lesen

durchgeführt werden. Im Anhang A.3 findet sich die tabellarische Beschreibung der Testhefte für die Domänen Deutsch Lesen (Tabelle A.4) und Englisch Leseverstehen (Tabelle A.5) äquivalent zur Tabelle 4.3 für Mathematik.

4.3.2. Simulierte Daten

Insbesondere die Untersuchung von IRT-Modellen erweist sich als ein beliebtes Feld für Simulationen, und so haben Feinberg und Rubright (2016) in ihrem Aufsatz Leitlinien für solche Simulationen entworfen. Dabei warnen sie davor, mit einer Simulation lediglich jene Annahmen zu „beweisen“, die Basis für die Erzeugung der Daten waren. Untersuchungen mit Hilfe von simulierten Daten werden von Ihnen deshalb eher als deduktive Beweise charakterisiert, selbst wenn diese nach einem experimentellen Design konstruiert sind (ebenda S. 37). Diese Argumentation geht davon aus, dass die Erzeugung von empirischen Daten im Rahmen einer Simulation keine zufällige Stichprobe der Realität erzeugt, sondern eine auf der Theorie begründete Datenbasis. Mit jedem simulierten Fall wird also die Theorie überprüft. Wenn Luecht und Ackerman (2018) schlussfolgern, dass simulierte Daten so „complicated and ‚messy‘ as the real data“ sein sollten, versuchen sie damit die Deutung von (bestimmten) Simulationen als induktives Vorgehen zu stärken. Die oben formulierten Hypothesen untersuchen allerdings die „pure Mechanik“ der Rasch-Skalierung und folgen damit, klassisch deduktiv, der Falsifikation theoretischer Annahmen.

Was ist mit „purer Mechanik“ gemeint? Die Simulation der Testbearbeitungen durch Personen mit definierter Personenfähigkeit und die auf der Basis dieser Testbearbeitung durch-

Tabelle 4.4.: Gegenüberstellung der Verteilung der Mathematikleistungen im Land Berlin bei VERA-8 2019 und der Leistungsverteilung der Simulation

		N	Mw[a]	Sd	Skew	Kurt	Quantil[a]				
							0.1	0.25	Median	0.75	0.9
ISS	Berlin	12762	0	110	-0.26	1.42	0	0	0	0	0
	Simulation	1250	+2	110	-0.11	-0.14	-14	-8	+7	+11	+2
GY	Berlin	11333	0	97	0.19	0.19	0	0	0	0	0
	Simulation	1250	-2	97	-0.07	-0.15	-12	-2	+8	+1	+6
Sum	Berlin	24095	0	125	-0.18	0.76	0	0	0	0	0
	Simulation	2500	+4	123	-0.20	-0.18	+2	-2	+11	+8	+4

[a]Hier werden lediglich die Unterschiede der Parameter aus der Simulation zu den Parametern aus dem Berliner Durchgang berichtet.

geführte Rasch-Skalierung bezieht keinerlei störenden Größen ein. So „bearbeiten“ die „Personen“ die Testaufgaben in vollständiger Konformität mit der Rasch-Theorie, niemand ermüdet während der Bearbeitung, niemand erinnert sich an eine sehr ähnliche Aufgabe der letzten Tage, niemand hat einen schlechten Tag, die Durchführungsbedingungen sind identisch bzw. irrelevant und es findet auch keinerlei fehleranfällige Bewertung offener Aufgaben statt. Es werden 100% Rasch-konforme Daten Rasch-Skaliert.

Simulation einer Population

Für die folgenden Untersuchungen wurde eine Verteilung simuliert[4], welche jener der Stadt Berlin ähnelt. Basis bildete die Vollerhebung der Mathematikleistungen im Rahmen von VERA 2019. Es wurden 2.500 Personenparameter simuliert, wobei jeweils die Hälfte einer Normalverteilung entspringt, deren Mittelwert und Standardabweichung denen aller Integrierten Sekundarschulen bzw. aller Gymnasien entspricht. In der Tabelle 4.4 sind die realen den simulierten Werten gegenübergestellt.

Für die Untersuchung der Fragestellungen ist es unerheblich, dass die Stichprobe die Situation Berlins gut abbildet, denn die Hypothesen beziehen sich auf jegliche Daten, müssen sich für jeden Datensatz als Gültigkeit erweisen. Durch eine Simulation, welche der Situation bei den Vergleichsarbeiten sehr ähnlich ist, lassen sich aber Auswirkungen von eventuell identifizierten Effekten auf die zwei schulformbezogenen Teilpopulationen wie auf die gesamte Verteilung einfach untersuchen und deren Wirkung für die reale Messung leichter interpretie-

[4]Zur Nachvollziehbarkeit wurden die Leistungswerte der Schüler*innen im R mit der Funktion *rnorm* unter Angabe der entsprechenden Parameter (wie in der Tabelle 4.4 gerundet) erzeugt. Der Zufallszahlengenerator wurde dazu über *set.seed* für die ISS-Schüler*innen mit dem Wert 323 und für die Gymnasiasten mit 454 voreingestellt. Die Simulation ist damit vollständig replizierbar.

ren.

Simulation von Itemantworten

Für diese 2.500 simulierten Personen mit ihren angemessen verteilten und bekannten Fähigkeitsparametern wurde für jedes Set an Itemparametern aus den 34 Testheften der letzten Jahre je eine Antwortmatrix simuliert. Die Syntax der R-Funktion (R Core Team, 2021) zur Erzeugung der Antwortmatrizen ist auf Seite 79 dokumentiert. Als Parameter werden mit der Itemschwierigkeit (item_pos = σ_i) sowie der Diskrimination der Items (item_dis = θ_i) zweierlei Itemparameter sowie die Personenparametern (pers = θ_v) übergeben. Zudem kann mit dem Parameter n_est angegeben werden, wie viele zufällige Replikationen jeder Antwortmatrix erzeugt werden sollen. Ohne Angabe wird genau eine Antwortmatrix simuliert. In der Funktion wird in Zeile 4 für jede mögliche Paarung von Personen-Fähigkeitsparametern und Item-Schwierigkeitsparametern die Differenz $\theta_v - \sigma_i$ gebildet. Diese entstehende Matrix wird in Zeile 6 mit der Diskrimination der Items δ_i gewichtet. Für eine Rasch-konforme Erzeugung wird für die Diskrimination ein 1-Vektor übergeben. Folgend werden in Zeile 9 über die Gleichung

$$P(X_{iv}) = \frac{e^{x_{iv}(\delta_i(\theta_v - \sigma_i))}}{1 + e^{\delta_i(\theta - \sigma_i)}} \tag{4.11}$$

die Wahrscheinlichkeiten für die Lösung jedes Items durch jede Person berechnet. Auf der Basis dieser Wahrscheinlichkeiten werden nun die Antworten simuliert. Dazu wird in Zeile 15/16 für jede Zelle der Matrix aus dem Tupel (0, 1) die 1 mit jeder dieser Wahrscheinlichkeiten gezogen; die 1 steht dabei für *richtig gelöst*, im anderen Fall die 0 für *falsch gelöst*. Mit dem Parameter n_est wird diese Zufallsziehung in einem Schritt mehrfach repliziert. Das von der Funktion zurückgegebene Objekt enthält die Matrix der Diskriminations-gewichteten Differenz von Item- und Personenparameter point_dis, die Wahrscheinlichkeiten für jede Person und jedes Item P und ein dreidimensionales Array der n_est simulierten Antwortmatrizen L.

R-Funktion zur Erzeugung von Zufallsantwortpattern (2PL-Modell)

```
1  reverse.rasch <- function(item_pos, item_dis, pers, n_est = 1) {

3  # calculate weighted distances a) distance (pers - item_pos)
4  point_pos <- outer(pers, item_pos, "-")
5  # b) weight (item_dis)
6  point_dis <- t(t(point_pos) * item_dis)
```

```
 8 # calculate probabilities
 9 p <- apply(point_dis, c(1, 2), function(x) exp(x)/(1 + exp(x)))

11 # random conditioned estimate of solution pattern
12 #   for every element of matrix make ...
13 #   take any number (n_est) random elements from c(0, 1)
14 #   with cover (T), with probability 1 - x for 0 and x for 1
15 k <- apply(p, c(1,2), function(x)
16                 as.logical(sample(c(0, 1), n_est, T, c(1 - x, x))))

18 # if n_est > 1 transpose the 3-dim-array in optimal form
19 if (n_est > 1) k <- aperm(k, perm = c(2, 3, 1))

21 # return list with 3 elements
22 reverse.rasch <- list(point_dis = point_dis, P = p, L = k)
23 }
```

Zahl der Replikationen

Die Zahl der Replikationen bei Simulationen gilt als Äquivalent zur Stichprobengröße bei Experimenten (Feinberg & Rubright, 2016). Mit einer größeren Zahl an Simulationen verringert sich der Standardfehler und entsprechend das Konfidenzintervall. Wie groß darf ein Konfidenzintervall sein, damit die Hypothesen mit ausreichender Sicherheit belegt werden können?

Für die Gültigkeit der Hypothesen 1 (S.65) und 2 (S.66) muss nachgewiesen werden, dass ein bestimmter Fähigkeitsschätzer im Ergebnis der Rasch-Skalierung bevorzugt wird und nicht der nächstgelegene. Die 3. Hypothese (S.71) zielt darauf ab, dass der Fähigkeitsschätzer zwischen zwei Itemparametern liegt. Diese können allerdings sehr dicht beieinanderliegen. Da die ermittelten Fähigkeitsschätzer nur an bestimmten Positionen liegen können, sollte aber auch hier ein bestimmter Fähigkeitsschätzer korrekt und dem nächstgelegenen vorgezogen werden. Offensichtlich darf das Konfidenzintervall um einen Fähigkeitsschätzer also nicht größer sein, als der halbe Abstand zum jeweils nächsten Item. Im linken Teil der Graphik 4.2 sind die Abstände der nach Schwierigkeit angeordneten Items für die 34 Testhefte dargestellt.

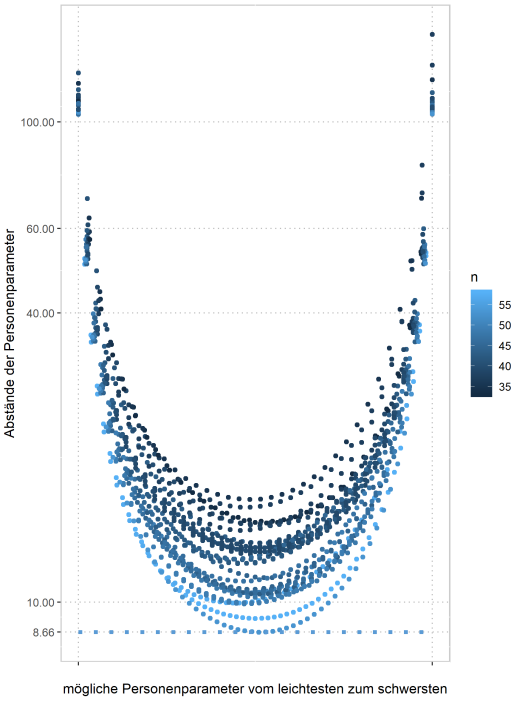


Abbildung 4.2.: Abstände der Personenparameter

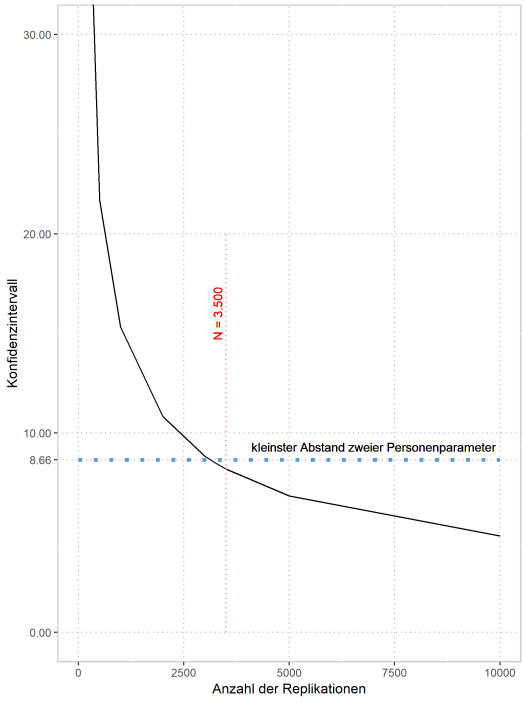


Abbildung 4.3.: Bestimmung der Replikationszahl

Jede „Perlenkette“ steht dabei für eines der 34 Testhefte. An der Blautönung ist die Zahl der Items der Testhefte grob ablesbar. Die Abstände um den Mittelwert der möglichen Personenparameter liegen am engsten beieinander. Der Mittelwert ist identisch mit dem Mittelwert der Itemparameter. Die Skala der Abstände auf der Ordinate ist logarithmisch abgetragen, damit die Differenzierung in der Mitte aber auch die großen Abstände an den Testhefträndern erkennbar bleiben. Den mit Abstand größten Abstand von über 100 Punkten haben die Personenparameter an den Rändern der Skala, also bei den sehr kleinen und den sehr großen Fähigkeitswerten. Das sind die approximierten Personenparameter für den Fall, dass Schüler*innen keine bzw. alle Aufgaben korrekt gelöst haben. Zur Mitte hin reduzieren sich die Abstände. Für Testhefte mit vielen Aufgaben (heller Blauton) rücken die Items in der Mitte dichter zusammen, als für Testhefte mit wenigen Aufgaben. Als minimaler Abstand über alle Testhefte und Items ergibt sich 8,6578.

Wenn das 95%-Konfidenzintervall maximal so groß sein darf, wie der kleinste Itemabstand, berechnet man aus der Standardabweichung der simulierten Population von 123,49 die minimale Anzahl von Replikationen zu

$$N_{rep} = \left(\frac{123.49}{\frac{8.6578}{2 \cdot 1.96}} \right)^2 = 3126 \tag{4.12}$$

Auf Basis der zur Verfügung stehenden Technik konnte die wünschenswerte Simulation mit 25.000 Schüler*innen und 3.500 Replikationen nicht durchgeführt werden. R hält alle Daten während der Berechnungen im Arbeitsspeicher, so dass dieser die gleichzeitig verarbeitbare Datenmenge begrenzt. Auf die 34 Testheft-Sets verteilen sich 1.428 Items. Für die 25.000 Schüler*innen ergeben sich so 35,7 Mio. Datenpunkte, bei 3.500 Replikationen knapp 125 Mrd. Deshalb wurde die simulierte Population mit 2.500 auf ein Zehntel reduziert. Da sich die 2.500 Fähigkeitsparameter letztendlich nur auf die möglichen Personenparameter verteilen (maximal 58 für das Set A von 2008), sollten für die Simulation ausreichend viele Personenparameter zur Verfügung stehen.

Mit dieser Reduktion der Personen und den folgend beschriebenen technischen Optimierungen, konnten 5.000 Replikationen realisiert werden. Es ergibt sich über den Standardfehler von

$$SE = \left(\frac{123.49}{\sqrt{5000}}\right) = 1.7464 \tag{4.13}$$

ein Konfidenzintervall von 6,8459, was deutlich unter dem kleinsten Abstand zweier Personenparameter liegt.

Technische Umsetzung der Simulation

Feinberg und Rubright (2016) wiesen darauf hin, dass in nur wenigen Veröffentlichungen die konkrete technische Umsetzung von größeren Replikationen thematisiert wird. Solche Hinweise sind allerdings für all jene, die sich solcher Methoden bedienen wollen, von essentieller Bedeutung. Deshalb wird im Folgenden kurz auf die technische Umsetzung eingegangen.

Die Berechnungen wurden anfangs auf einem leistungsfähigen Desktop-System durchgeführt, welches für 2.000 Replikationen knappe 16 Stunden benötigte (siehe Tabelle 4.5). Die für noch längere Berechnungen notwendige Betriebsumgebung konnte auch wegen der Administration durch das Rechenzentrum (nicht kontrollierbare Wartungsintervalle, automatischer Übergang in den Standby-Modus über Nacht, etc.) nicht sichergestellt werden. Deshalb wurden die Berechnungen letztendlich auf einen leistungsmäßig äquivalenten virtuellen Server ausgelagert. Auf dem Linux-Server war R-Studio als serverseitige Webapplikation installiert. R wurde hier remote betrieben, so dass die Berechnungen autark laufen konnten. Durch die Optimierung der Prozeduren konnten hier 5.000 Replikationen sicher realisiert werden.

Grundsätzlich erwiesen sich zwei Datenstrukturen als „Speicherfresser“: (a) die abgespeicherten Ergebnisse, wobei sämtliche geschätzten Personenparameter gespeichert wurden und (b) die temporär erzeugten Antwortmatrizen. Die Speicherung der Antwortmatrizen wurde

Tabelle 4.5.: Rechenzeiten unterschiedlicher Systemen im Vergleich

Desktop-PC	virtueller Server
Intel (R) Core i7-6700 @ 3,40GHz mit 4 physischen und 8 logischern Kernen RAM: 16GB	Intel (R) Xeon (R) Gold 5120 @ 2,20GHz mit 8 logischen Kernen (genutzt: einer) RAM: 16GB
Rechenzeit für 2000 Replikationen: 15h	Rechenzeit für 2000 Replikationen: 16h Rechenzeit für 3500 Replikationen: 25,5h Rechenzeit für 5000 Replikationen: 36h

in einer ersten Optimierung auf das Speicherformat auf LOGICAL[5] reduziert, was die Speicherauslastung durch dieses Objekt von 2,1 GB um ca. 1 GB reduzierte. Das bedeutet aber auch, dass für eine Simulation mit deutlich größeren Populationen als 5000 ein größerer Arbeitsspeicher bzw. eine Umstellung der Programmierung notwendig wäre. Zudem wurden die Zwischenergebnisse für jede Kombination von Jahr, Version und Diskrimination auf der Festplatte gespeichert und als Objekt temporär gelöscht. Die Daten wurden dann später wieder zusammengefügt.

Abbildung 4.4 zeigt die Auslastung der Serverressourcen für die gut 35 Stunden andauernde Simulation mit 5.000 Replikationen auf einem virtuellen Server. Es wird deutlich, dass für eine größere Anzahl von Replikationen (oder Personen) der verfügbare Arbeitsspeicher angepasst werden müsste.

Im Zeitraum der Berechnungen von 13:49 bis 00:58 ist die CPU nur gering ausgelastet. Das liegt daran, dass R standardmäßig nur einen Kern „beschäftigt“. Damit R mehrere Kerne verwendet, müsste der R-Job entsprechend umgebaut werden, was allerdings lediglich die Dauer des Jobs reduzieren würde. Da jeder parallel arbeitende Kern eigenen Arbeitsspeicher benötigt, stiege der Bedarf hier proportional an. Eine Optimierung des Codes bezüglich der Nutzung mehrerer Kerne lohnte sich also nur dann, wenn die Optimierung mindestens in gleichem Maße auch den Bedarf an Arbeitsspeicher reduzieren würde.

4.4. Ergebnisse

Im Folgenden werden die Ergebnisse der Überprüfung der vier zuvor aufgestellten Hypothesen zur Rasch-Skalierung vorgestellt.

[5]Der Datentyp LOGICAL unterscheidet lediglich TRUE und FALSE und benötigt somit theoretisch nur 1 bit, der sonst verwendete Typ INTEGER im Allgemeinen 1 Byte, also 8 Mal so viel. Die tatsächliche Reduktion betrug allerdings nur ca. 50%.

Abbildung 4.4.: Auslastung von CPU (oben) und RAM für 3500 Personen und 5000 Replikationen

4.4.1. Irrelevanz der Itemauswahl

Hypothese 1: Unabhängig von der Auswahl der Items für ein Testheft, ergeben sich für eine Person nur zufällig verschiedene, im Mittel identische Fähigkeitsschätzer.

In der ersten der zu überprüfenden Hypothese wird formuliert, dass es im Rahmen der Rasch-Skalierung und basierend auf einer Menge von Items mit bekannten Schwierigkeiten gleichgültig für die Schätzung der Fähigkeiten ist, mit welcher Itemauswahl die Schätzung vorgenommen wird. Für die Prüfung dieser Hypothese wird

- die beschriebene simulierte Verteilung von Personenfähigkeiten verwendet sowie
- die 34 Sets von Items aus allen vergangenen VERA-8-Durchgängen.

Aus der Abbildung 4.5 über die mittleren Schwierigkeiten der Testhefte über die letzten Jahre lässt sich entnehmen, dass die vornehmlich für nicht-Gymnasien bestimmten Testheftversionen 1 und 2 über die Jahre deutliche Anpassungen erfahren haben. Für das Testheft 1 liegt die mittlere Schwierigkeiten zwischen 334 (2013) und 461 BiSta-Punkten (2010). Das schwierigste Testheft, vom IQB als Testheft für Gymnasien konzipiert, weist demgegenüber eine relativ konstante mittlere Schwierigkeit auf. Andererseits lagen die Schwierigkeiten der drei Testhefte 2010 mit 72 Punkten deutlich enger beisammen, als 2012, wo ein Bereich von 198

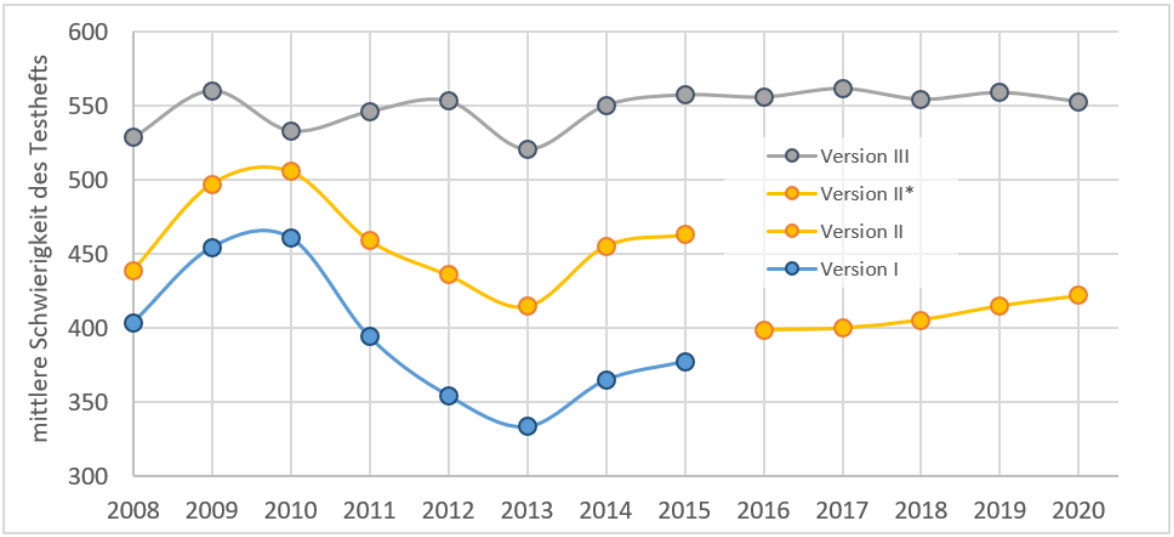


Abbildung 4.5.: Mittlere Schwierigkeit der Items für die Testheftversionen bei VERA-8 Mathematik von 2008 bis 2020

Punkten überstrichen wird. In den letzten Jahren bis 2015 hat sich für die drei Versionen ein Abstand von gut 80 bis 100 BiSta-Punkten zwischen Testheft 1 und 2 sowie zwischen 2 und 3 eingestellt. Die aktuell zwei Testheftversionen liegen seit 2016 recht stabil bei rund 400 und 550 BiSta-Punkten. Für die zwei Domänen Deutsch Lesen und Englisch Leseverstehen sind die Verhältnisse etwas anders, aber strukturell ähnlich, wie man den Abbildungen A.1 und A.2 im Anhang entnehmen kann.

Damit stellen die 34 Testhefte bezüglich ihrer Schwierigkeit sehr unterschiedliche Zusammenstellungen von Items dar. Geht man davon aus, dass ein Testheft gut zur Population passt, wenn die mittlere Schwierigkeit der Items in etwa der mittleren Fähigkeit der Schüler*innen entspricht und diese zwischen den Jahren nur moderat schwankt, passen die verschiedenen Testhefte unterschiedlich gut. Um die Güte der Schätzung der Fähigkeit θ durch die Rasch-Skalierung für die 2.500 Schüler zu ermitteln, wurde mit dem Bias die Summe der Differenz zwischen der Schätzung und dem wahren Wert über alle $n = 5.000$ Replikationen

$$Bias = \frac{\sum_{i=1}^{n}(\theta_i - \theta_{True})}{n} \tag{4.14}$$

und darüber hinaus der Standardfehler

$$sd_{Bias} = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \theta_{True})^2}{n-1}} \tag{4.15}$$

ermittelt.

Simulation ohne Berücksichtigung der Diskrimination

Die Abbildung 4.6 zeigt die Verteilung der Messfehler und den Standardfehler beispielhaft für das einfache (oben) und das schwere Testheft (unten) des Jahres 2015. Auf der Abszisse ist die Skala der Bildungsstandards abgetragen. Für das jeweilige Testheft sind in der Graphik unterhalb der Nulllinie die Items mit Ihrer Schwierigkeit (o) und oberhalb der Nulllinie die resultierenden möglichen Personenparameter (Δ) in Bezug zur waagerechten Skala der Bildungsstandards eingetragen. Jede mit einem roten Punkt simulierte Personenfähigkeit wird im Rahmen einer Rasch-Skalierung einem der möglichen Personenparameter (Δ) zugeordnet. Wegen der mehrfachen Replikation der Skalierung, wird hier der Mittelwert dieser Zuordnungen abgebildet. Um dies beispielhaft zu verdeutlichen: Ist eine Personenfähigkeit von 300 Punkten simuliert worden, dann ordnet die Rasch-Skalierung dieser Person einen der möglichen Personenparameter zu. Dieser mögliche Personenparameter liegt dabei in der Nähe des wahren Parameters von 300, aber nicht notwendig beim nächstgelegenen, sondern auch mal zwei, drei, vier mögliche Personenparameter daneben. Im besten Fall sollte der Mittelwert der vielfach replizierten Zuordnungen bei 300 liegen. Auf der Ordinate ist für jede simulierte Personenfähigkeit die Differenzen aus der Schätzung des Personenparameters (also dem Mittelwert der mehrfach replizierten Zuordnung) und dem bekannten wahren Personenparameter abgetragen, also der Bias nach Formel 4.14. Der rote Korridor zeigt für jeden Wert die Streuung des Bias auf der Basis der Replikationen an (Formel 4.15).

Im Wesentlichen findet sich erwartungsgemäß nur eine zufällige Streuung um die Nulllinie. Um den Mittelwert der Itemschwierigkeiten[6] ist der Standardfehler am kleinsten und er wächst zu den Rändern hin erst moderat und später deutlicher an. Lediglich im oberen Bereich des einfachen Testhefts und etwas stärker noch im unteren Bereich des schwierigen Testhefts sind Verzerrungen zu erkennen. Woher kommen diese Verzerrungen? Die Verzerrungen werden dort sichtbar, wo auf Grund der Lage der Items nur noch wenige mögliche Personenparameter existieren, wohl aber Personen mit einer Fähigkeit. In der oberen Graphik liegen die letzten drei möglichen Personenparameter am unteren Skalenrand bei unter -50 Punkten, bei etwas über 0 Punkten und bei ca. 50 Punkten. Da die Person mit dem kleinsten simulierten Fähigkeitswert mit etwa 70 Punkten darüber liegt, wird diesen drei untersten Personenparametern ein Anteil der Replikationen zugeordnet, wie auch einigen darüber liegenden möglichen Personenfähigkeiten. Im Bereich des letzten und vorletzten Personenpa-

[6] Der Mittelwert der Itemschwierigkeiten und der Mittelwert der möglichen Personenparameter sind identisch. Die Lage ist in der Graphik durch das graue Dreieck markiert.

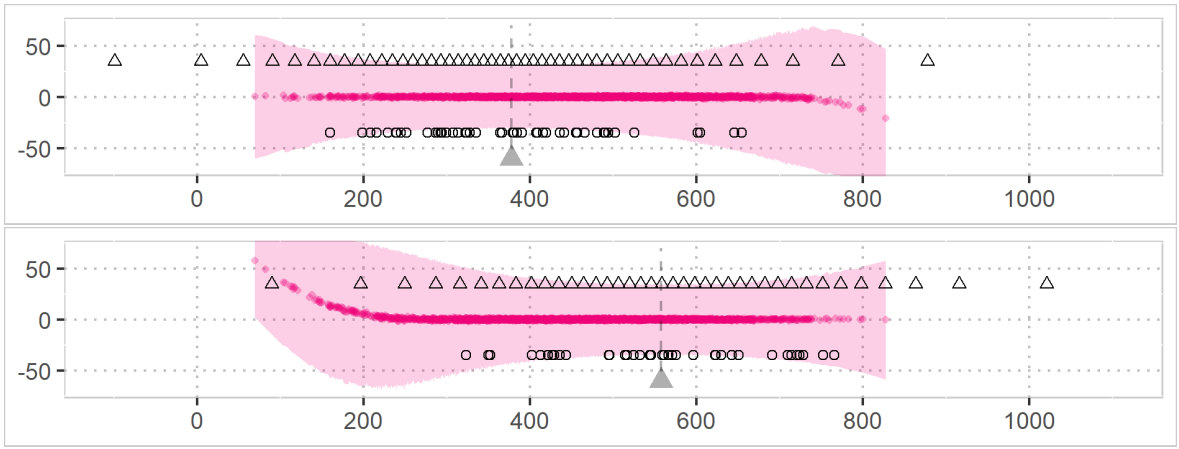


Abbildung 4.6.: Differenz von wahrer und geschätzter Fähigkeit (Bias, inkl. dessen Standardabweichung) vor dem Hintergrund der Verteilung der Item- (o) und Personenparameter (Δ) für das leichte (oben) und das schwierige Testheft (unten), beispielhaft für das Jahr 2015.

rameters am oberen Skalenrand finden sich allerdings einige Personen mit hohen simulierten Fähigkeiten. Auch diese werden nun im Rahmen der Rasch-Skalierung möglichen Personenparametern zugeordnet. Für die Person mit einer simulierten Fähigkeit von 800 kann aber oberhalb nur ein einziger möglicher Parameter gefunden werden, unterhalb deutlich mehr. Deshalb werden Personen mit einer Fähigkeit von 800 bei etwa symmetrischer Streuung eher niedrigeren möglichen Personenparametern zugeordnet als dem einen höheren. Im Ergebnis der Mittelung zeigt sich hier deshalb eine tendezielle Unterschätzung. Diese Unterschätzung trifft notwendig auf alle Personen zu, deren Fähigkeit die des höchstmöglichen Personenparameters übersteigt. Das ist quasi ein Deckeneffekt des leichten Testhefts für Schüler*innen mit sehr hohen Leistungen. Äquivalent findet man in der Abbildung 4.6 (unten) den Bodeneffekt für Schüler*innen mit eher schlechter Performance, denen das schwere Testheft vorgelegt wird. In der Abbildung 4.9 erkennt man diese Verzerrung deutlich daran, dass die Schar der einem bestimmten möglichen Personenparameter zugeordneten wahren Werte an den Rändern nicht mehr symmetrisch um den Personenparameter verteilt ist. Auf der anderen Seite der Skala findet sich das inverse Phänomen. Natürlich ist bekannt, dass das Konfidenzintervall an den Rändern ohnehin größer wird (Rost, 2004, S.356 ff). Hier aber addiert sich eine vielleicht erwartbare zusätzliche Verzerrung.

Diese Darstellung findet sich für alle 34 Testhefte in äquivalenter Form (siehe die margentafarbenen Punktwolken in der Abbildung A.3 im Anhang). Grundlage dieser Analyse sind simulierte Daten. Alle Schüler*innen verhalten sich Rasch-konform. Die Fit-Statistiken für Infit und Outfit liegen erwartungsgemäß mit ±0.03 sehr eng um den idealen Wert von 1. So

perfekt passende Items sind in realen Untersuchungen kaum anzutreffen. Trotzdem finden sich Verzerrungen dort, wo Schüler*innen bezüglich der Schwierigkeit schlecht angepasste Testhefte vorgelegt wurden.

Simulation mit Berücksichtigung der Diskrimination

Schließlich muss festgestellt werden, dass die bei der Rasch-Skalierung nicht modellierte Diskrimination (oder auch Trennschäfte) in realen Untersuchungen von der für die Rasch-Skalierung bei 1 fixierten abweicht (siehe Spalte „Diskrimination Mw“ in Tabelle 4.3), d.h. die reale Itemcharakteristik weist eine von 1 differente Diskrimination auf. Diese Abweichungen liegen zudem deutlich häufiger in einem Bereich größer 1 als kleiner 1. Das bedeutet, die Items sind trennschärfer, als das für das Rasch-Modell erwartet wird. Dass entgegen der erwarteten Streuung um 1 vermehrt trennscharfe Items zu finden sind, lässt sich zumindest teilweise aus den Ausführungen zur Itemselektion in den technischen Berichten zu VERA erklären. Die folgende Formulierung zur Itemauswahl (beispielhaft entnommen aus Aneis et al., 2018) auf Basis des Infits, ersetzt seit 2012 eine in den zwei Jahren zuvor äquivalent formulierte.

> „Der Infit sollte nicht größer als 1,10 ausfallen, wobei Werte bis 1,15 in Ausnahmefällen toleriert wurden. Eine Untergrenze wurde nicht definiert, da Items mit einem Infit kleiner 1 zwar keine optimale Passung zum Raschmodell aufweisen, aber trennschärfer als modellkonform sind, was eine positive Eigenschaft darstellt.“ (ebenda, S.12)

Eine weniger positive, bezüglich der Konsequenzen aber ähnliche Einschätzung, nach der solche Items für die Messung zwar wenig nützlich sind, die Messung aber auch nicht verschlechtern, lediglich ggf. zu irreführend guten Reabilitäten führen (Wright & Linacre, 1994), wurde später mit dem Hinweis ergänzt: „If we are developing a new test: replace overfitting items with more efficient items.“[7].

Um den Einfluss dieser zugelassenen Variation zu untersuchen, wurde die zufällige Simulation von Antwortmatrizen der Schüler*innen dahingehend variiert, dass neben der Raschkonformen Diskrimination von 1,0, auch solche von 0,7 sowie 1,7 herangezogen wurden. Um den Effekt deutlich separieren zu können, wurden in jeweils einem Durchgang alle Items mit der gleichen Diskrimination belegt. Die Berechnung der Personenparameter mit Hilfe der

[7] Die Quelle verweist auf eine Website, auf der die Ergänzung lediglich mit *Later* eingeleitet wurde, allerdings nicht ersichtlich wird, wann diese Ergänzung vorgenommen wurde.

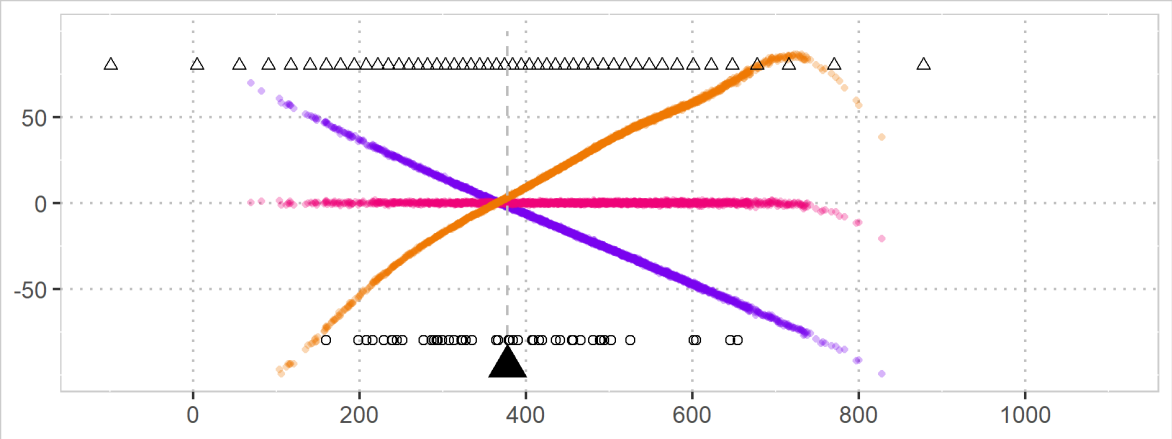


Abbildung 4.7.: Differenz von wahrer und geschätzter Fähigkeit für das leichte Testheft (2015) für den Fall, dass die Personen die Items mit einer Diskrimination von 0,7 (violett), 1,0 (pink) bzw. 1,7 (gelb) bearbeiten.

Rasch-Skalierung und die 5.000-fache Replikation blieb davon unberührt.

Die Abbildungen 4.7 und 4.8 finden sich in genau dieser Form für sämtliche 34 Testhefte im Anhang (vergleiche Abbildung A.3): Mit größer werdender Differenz von mittlerer Testheftschwierigkeit (mit einem schwarzen Dreieck in der Abbildung gekennzeichnet) und Personenfähigkeit führen von 1 verschiedene Diskriminationen zu einer Verzerrung der Schätzung der Personenparameter. Bekommt eine Person ein unangemessen leichtes Testheft, dessen Items zudem eine durchschnittliche Diskrimination von größer 1 (gelbe Punkte) aufweisen, so wird dessen Fähigkeit unterschätzt. Für eine gegebene Diskrimination zeigt sich eine im mittleren Bereich etwa lineare Abhängigkeit der Verzerrung der Messung der Personenfähigkeit von der Differenz zwischen der mittleren Personenfähigkeit und der mittleren Itemschwierigkeit. Lediglich die zuvor schon festgestellten Verzerrungen wegen fehlender Aufgaben am Skalenrand überlagern diesen Effekt und dämpfen damit diese Verzerrung ein wenig (insbesondere zu erkennen in Abbildung 4.8).

Zu erkennen ist allerdings auch eine starke Abhängigkeit der Verzerrung von der konkreten Lage der Items. Als besonders auffälliges Beispiel zeigt die Abbildung 4.8 die Verzerrungen für das Testheft C des Jahres 2009. Die Hypothese 1, nach der die Auswahl von Items irrelevant ist und der Einsatz unterschiedlicher Testhefte damit zu identischen Ergebnissen führt, kann nur unter der Bedingung aufrechterhalten werden, dass die Diskriminationen der Items nahe 1 liegen oder dass Testhefte immer eine enge Passung zu den Schülerfähigkeiten aufweisen[8].

In realen Daten variieren die Diskriminationen einzelner Items natürlich. Eine Übersicht

[8]Diese enge Passung ist natürlich nicht herstellbar, weil man dann vor der Messung schon die zu messende Personenfähigkeit kennen müsste.

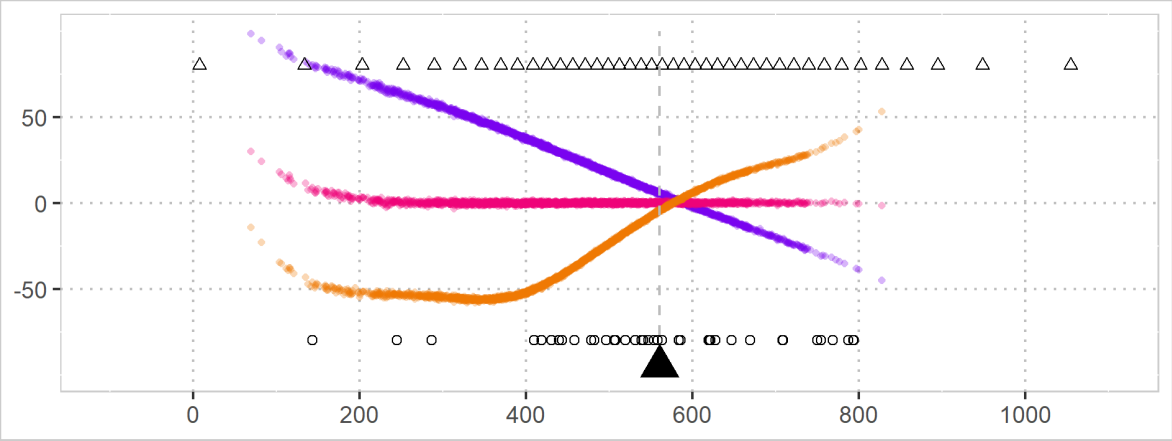


Abbildung 4.8.: Differenz von wahrer und geschätzter Fähigkeit für das schwierige Testheft (2009) für den Fall, dass die Personen die Items mit einer Diskrimination von 0,7 (violett), 1,0 (pink) bzw. 1,7 (gelb) bearbeiten.

zu den mittleren Diskriminationen aus den Erhebungen in Berlin wird auf die an angemessen großen Stichproben (N > 250) eingesetzten Testhefte beschränkt. Nicht berücksichtigt wurde damit der Einsatz der Testhefte der Versionen C an kleinen Stichproben ausgewählter Klassen, wie das Mathematik-Testheft in Mathe-Profilklassen bzw. die Fremdsprachen-Hefte in bilingual unterrichteten Klassen. Für die 26 in Berlin eingesetzten Testhefte wurde für die Simulation der Antworten für jedes Item nicht nur die aus der Pilotierung durch das IQB bekannte Schwierigkeit, sondern auch die aus dem Einsatz in Berlin ermittelte itemweise Diskrimination berücksichtigt. Die Diskriminationen wurden im Rahmen von Skalierungen der Berliner Vollerhebungen nach dem 2-PL-Modell (auch Birnbaum-Modell) geschätzt, wobei die Schwierigkeiten der Items auf die vom IQB ermittelten Parameter aus der Pilotierung fixiert wurden. Die Schätzungen der Personenparameter erfolgten in der Simulation allerdings weiterhin nach dem Rasch-Modell. Wieder wurde die gleiche, Berlin nachempfundene Population herangezogen und die Simulation der Antworten wurde wieder 5.000-fach repliziert.

Über die Darstellungen der 26 eingesetzten Testhefte hinweg (vergleiche die Abbildung A.4 im Anhang A.5) ist wieder zu erkennen, dass im Punkt der mittleren Itemschwierigkeit keine relevante Verzerrung vorzufinden ist. Die Punktwolke der 2.500 Fähigkeitsschätzungen schneidet in jedem Fall in genau diesem Punkt die Bias-Nulllinie. Abseits davon ist der Verlauf der Verzerrung durch die vorherige Simulationen mit festen Diskriminationen vorgezeichnet, nun aber doch individuell. Die Verzerrungen sind wie erwartet kleiner aber auch individueller. In der vorhergehenden Simulation wurden die Diskriminationen aller Items auf 0,7 bzw. 1,7 gesetzt. Tatsächlich liegt der Mittelwert der Diskrimination bei den eingesetzten Testheften zwischen 0,93 und 1,48 und weicht damit deutlich weniger von 1 ab. Deshalb sind die Verzer-

rungen eher kleiner als in der ersten Simulation mit festen Diskriminationen. Zum anderen hängen die Verzerrungen in der Realität deutlicher mit den unterschiedlichen Diskriminationen der verschiedenen Items und deren Positionen zum jeweils geschätzten Personenparameter zusammen, wie das schon in der Abbildung 4.8 deutlich wurde. Das bedeutet konkret, dass Verzerrungen für eine Person größer werden können, wenn deren Personenparameter in der Nähe von Items liegt, die in besonderem Maße eine von 1 abweichende Diskrimination aufweisen.

Unter den gegebenen Umständen speziell der im technischen Manual des IQB (Aneis et al., 2018) ausgeführten Itemselektion, muss die *Hypothese 1* für die Vergleichsarbeiten verworfen werden.

4.4.2. Erwartungstreue Schätzung von Personenparametern

Hypothese 2: Der Mittelwert der wahren Personenparameter aller durch die Rasch-Skalierung einem Personenparameter zugeordneten Personen liegt in direkter Nähe dieses Personenparameters. Es gibt keinen zweiten möglichen Personenparameter, der dichter liegt.

Zur Überprüfung dieser Hypothese wird wieder auf ein beliebiges Set von Itemparametern und eine normalverteilte Fähigkeitsverteilung zurückgegriffen. Schülerinnen und Schüler mit identischer Anzahl richtig gelöster Items wird der gleiche Personenparameter zugeordnet. Die Darstellung 4.9 verbindet für alle Schüler*innen (N = 2.491) den simulierten wahren Fähigkeitswert (oben) mit den aus der Rasch-Skalierung geschätzten (unten). Die Zuordnungen zu den zwei extrapolierten Werten für Schüler*innen die keine bzw. alle Aufgaben gelöst haben (N = 9), wurden hier nicht betrachtet. Die Schar die von jedem fünften Schätzwert ausging wurde zur besseren Übersichtlichkeit gelb, ein mittlerer und ein oberer Bereich noch einmal rot hervorgehoben.

Die Darstellung bezieht sich wieder auf den Einsatz des recht durchschnittlich auffälligen Mathematik-Test des Jahres 2015 mit der leichten Version A (also an nicht-Gymnasien Berlins). Zwei Aspekte fallen dabei ins Auge, betrachtet man jeweils die Schar der Zuordnungen zu einem Schätzwert:

1. Die wahren Werte streuen in einem weiten Bereich um den Schätzwert. Nur für wenige der Zuordnungen trifft es offensichtlich zu, dass sie dem nächstliegenden möglichen Personenparameter zugeordnet sind.

Tatsächlich liegt im Beispiel für innere Personenparameter (+/- eine Standardabweichung von 202 um den Mittelwert von 378) die Streuung der zugeordneten wahren Werte zwischen

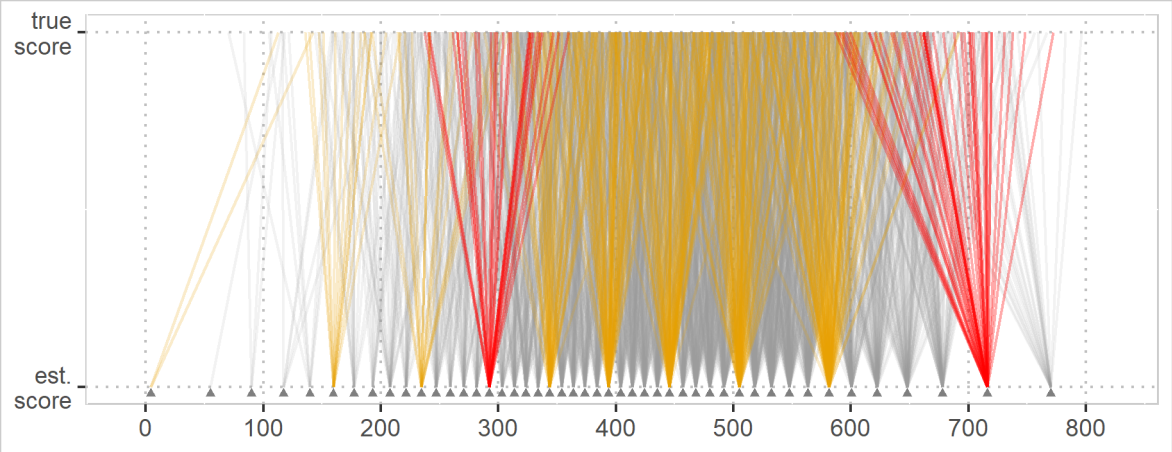


Abbildung 4.9.: Gegenüberstellung der (simulierten) wahren Fähigkeit (oben) und der durch die Rasch-Skalierung geschätzten Fähigkeit (unten), beispielhaft für das Testheft A, 2015.[9]

27 und 37 Skalenpunkten und erreicht an den Randbereichen Werte um 60. Diese Streuung ist damit fast identisch mit dem Standardfehler, wie er sich bei der Rasch-Skalierung für jeden Personenparameter ergibt. Die Darstellung ist somit ein bildhafter Ausdruck für das Konfidenzintervall bei der Rasch-Skalierung. Weiter vorn wurde schon festgestellt, dass die Personenparameter nur mit relativ großer Unsicherheit geschätzt werden können, weil dazu nur die wenigen Itemparameter zur Verfügung stehen. Diese große Streuung war demnach zu erwarten.

2. Während in der Mitte der Verteilung die Linien von einem Schätzwert (unten) aus symmetrisch nach links und rechts verteilt zu sein scheinen, trifft dies auf die Randbereiche nicht zu.

Die Abbildung 4.10 zeigt die gleichen Werte des obigen Beispiels in anderer Form: Die Graphik stellt auf der auf der Abszisse die BiSta-Skala dar, wobei nur für die Punkte der möglichen Personenparameter Werte abgetragen werden. Diese Punkte sind durch kleine Dreiecke am unteren Rand gekennzeichnet. Alle Werte sind auf die für diese Punkte geschätzten Werte bezogen, welche die Nulllinie darstellen. Mit dem roten Band sind für jeden geschätzten Wert die halben Abstände zum vorhergehenden und nachfolgenden möglichen Personenparameter abgetragen. Dafür wurde vom aktuellen Wert der des jeweiligen Nachbarn subtrahiert und die Differenz halbiert. Der linke Nachbar ist daher im roten Band nach oben abgetragen, der rechte Nachbar nach unten. Da die Abstände zu den Rändern hin zunehmend größer werden, ist das rote Band im unteren Bereich nach oben und im oberen Bereich nach unten breiter. Dargestellt sind allerdings nur die mittleren 47 der durch die Rasch-Skalierung berechneten

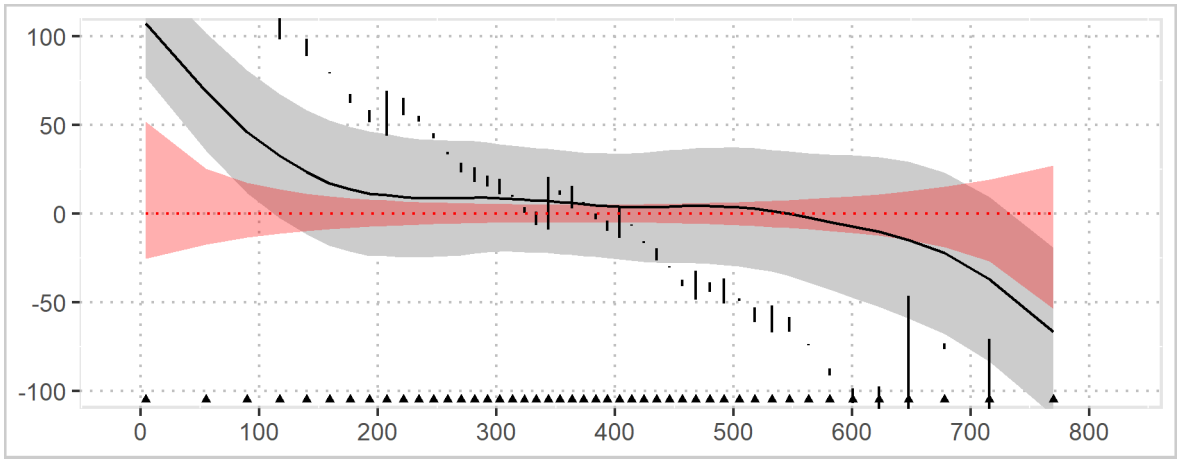


Abbildung 4.10.: Alle dargestellten Werte sind auf die geschätzten Personenparameter bezogen, welche die Nulllinie bilden. Die Punkte der möglichen Personenparameter auf der BiSta-Skala sind am unteren Rand der Darstellung durch kleine Dreiecke gekennzeichnet. Der rote Bereich beschreibt die halben Abstände zu den nächstliegenden Schätzwerten links (oben) und rechts (unten). Die jedem möglichen Personenparameter zugeordneten wahren Werte sollten vom Schätzwert nicht mehr abweichen, als dieser rote Bereich. Die schwarze Linie verbindet die Mittelwerte all jener wahren Werte, welche diesem Personenparameter zugeordnet wurden. Der graue Bereich beschreibt eine Standardabweichung dieser Mittelwerte.

49 Personenparameter, die äußeren zwei sind, wie oben schon dargestellt, lediglich durch eine Extrapolation berechnet worden. Für die Darstellung wurden die aus den 3.500 Rasch-Skalierungen für eine konkrete Simulation berechneten 2.500 Personenparameter verwendet. Von den 8,75 Millionen Personenparametern entfielen 0,36% auf die zwei Personenparameter, die keine oder alle Aufgaben korrekt gelöst haben und hier nicht dargestellt werden, so dass diese Daten auf gut 8,7 Millionen Personenparametern beruhen. Auf jeden der 47 möglichen Personenparameter entfällt nun ein bestimmter Anteil der 2.500 wahren Personenfähigkeiten. Für alle jeweils genau einem möglichen Personenparameter zufallenden wahren Personenfähigkeiten wird der Mittelwert und die Standardabweichung berechnet und als schwarze Linie bzw. graues Band verbunden dargestellt. Immer dort, wo die schwarze Linie der wahren Mittelwerte im roten Bereich des halben Abstands ist, liegt der Mittelwert der wahren Werte dem geschätzten Wert am nächsten, im anderen Fall einem weiter entfernt liegenden Schätzwert.

Das folgende Lesebeispiel (siehe Beispiel Abbildung 4.11 verdeutlicht die Bedeutung der Graphik 4.10 für einen bestimmten Punkt und zwar für den neunten hier abgebildete Personenparameter bei 207,63 BiSta-Punkten. Es ist der 10. Personenparameter in der Reihenfolge und damit derjenige, der einer Person zugeordnet wird, die 9 Aufgaben korrekt lösen konnte. Der halbe Abstand zum vorhergehenden Personenparameter 192,88 beträgt 7,37 und zum

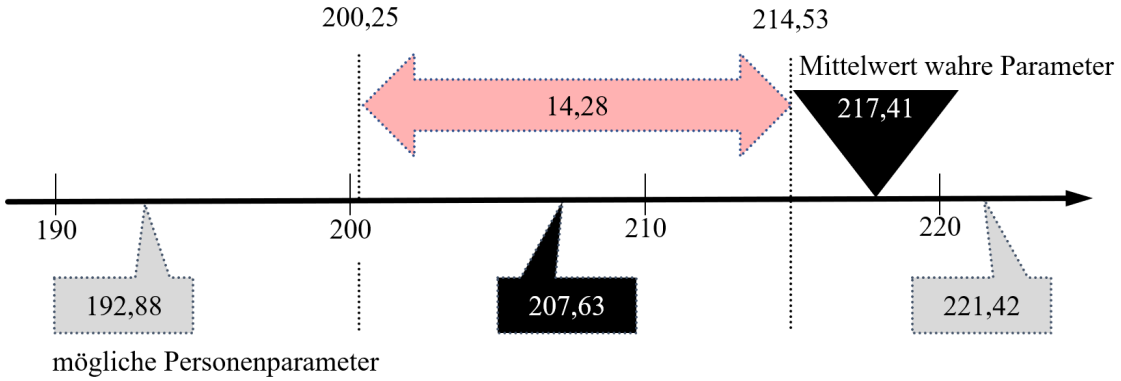


Abbildung 4.11.: Lesebeispiel für den Punkt bei 208 BiSta-Punkten aus der Abbildung 4.10. Der Mittelwert für alle wahren Parameter, die dem Personenparameter bei 207,63 zugeordnet wurden, liegt bei 217,41 und damit außerhalb des halben Abstandes zum nächstliegenden Parameter bei 221,42.

nachfolgenden Parameter 221,42 beträgt 6,90 Punkte. Dieser um den Ausgangspunkt 207,63 nicht symmetrische Bereich, ist in beiden Abbildungen als rotes Band dargestellt. Der Mittelwert der wahren Werte, die diesem Personenparameter zugeordnet sind liegt mit 217,41 außerhalb dieses Bereiches. Er ist in der Abbildung 4.10 als schwarze Linie und im Lesebeispiel der Abbildung 4.11 mit einem schwarzen Dreieck dargestellt. Wie hier liegt für viele Schätzungen der Mittelwert der wahren Werte nicht dem zugeordneten Schätzwert am nächsten. Im vorliegenden Fall trifft das lediglich auf die 20 Personenparameter im Bereich vom 25. (383,65) bis zum 44. Personenparameter (647,66) zu und nicht für die 29 anderen Parameter.

Oft werden Verzerrungen im Rahmen einer Rasch-Skalierung begleitenden Umständen der Messung zugeschrieben. In der hier vorliegenden Untersuchung, dass sei nochmal herausgehoben, liegen keine solchen schwer konstant zu haltenden oder gänzlich zu verhindernden Einflüsse einer üblichen Testung vor. Die Analysen geschehen auf der Basis simulierter Personenfähigkeiten, die in hochgradig Rasch-konformer Weise Fragen „bearbeiten“, deren Schwierigkeit an genau jenen Punkten liegen, wie sie sich in den 34 Testheften der Vergleichsarbeiten finden. Hier hätte natürlich auch jede andere Verteilung von Itemschwierigkeiten verwendet werden können. Diese Verzerrung ist demnach nur auf Effekte zurückzuführen, die der Rasch-Skalierung selbst eingeschrieben bzw., wie weiter oben benannt, Ergebnis der *puren Mechanik* der Rasch-Skalierung sind.

Weder für jeden einzelnen wahren Personenparameter (siehe Abbildung 4.9) noch für den Mittelwert aller wahren Personenparameter, die einem geschätzten Personenparameter zugeordnet werden (siehe Abbildung 4.10), trifft zu, dass diese immer in direkter Nähe des geschätzten Parameters liegen. Die *Hypothese 2* muss somit verworfen werden.

4.4.3. Die Bedeutung von Gutmann-Pattern

Hypothese 3: Eine mit Hilfe der Rasch-Skalierung definierte Metrik, ist gegenüber Guttman-Pattern unverzerrt.

Für die Überprüfung der Hypothese 3 wird wieder zuerst exemplarisch auf die A-Version des Mathematik-Testhefts des Jahres 2015 zurückgegriffen, die Betrachtungen später aber wieder auf alle Testhefte übertragen. Der einem Guttman-Pattern zugeordnete Personenparameter gilt dann als unverzerrt, wenn er zwischen dem letzten richtig gelösten und dem ersten nicht richtig gelösten Item liegt. In der Abbildung 4.10 beschreibt jeweils ein schwarzer Balken den Bereich zwischen genau den zwei Itemparametern, zwischen denen der Personenparameter für das zugeordnete Guttman-Pattern liegen sollte, wenn dieser unverzerrt aus der Rasch-Skalierung hervorgehen soll.

Für das Beispiel des zehnten möglichen Personenparameters für neun korrekt gelöste Aufgaben bei 207,63 würde erwartet, dass dieser zwischen dem neunten und zehnten Itemparameter liegt, also zwischen 251,29 und 276,81. Dieser Abstand wurde in der Abbildung 4.10 wieder bezogen auf den geschätzten zugeordneten Personenparameter bei 207,63 als schwarzer Balken eingezeichnet, der hier von 251,29 - 207,63 = 43,66 bis 276,81 - 207,63 = 69,18 verläuft. Der Balken ist kleiner, wenn zwei der Schwierigkeit nach geordneten, aufeinander folgenden Items dicht beieinander liegen. Schneidet der Balken die Nulllinie, bedeutet dies, dass der geschätzte Personenparameter (Nulllinie) im Intervall zwischen den zwei benachbarten Items (schwarzer Balken) liegt. Für diese Personenparameter ist die Hypothese zutreffend. Beim der Graphik zugrundeliegenden Testheft finden sich nur in der Mitte der Verteilung genau vier Personenparameter, für die dies zutrifft. Hier liegen, zu erkennen an den kurzen schwarzen Balken, die Items teilweise sehr dicht beieinander. In der Regel liegen die Parameter von Personen, deren Antwort ein Gutmann-Pattern darstellt, nicht zwischen den zwei Items, welche den Übergang von gelösten zu ungelösten Items beschreiben. Dass die Balken im unteren Bereich der BiSta-Skala nach oben abweichen und im oberen Bereich umgekehrt ist natürlich mit dem Umstand verbunden, dass die möglichen Personenparameter deutlich breiter verteilt sind, als die Itemparameter. Allerdings trifft die Annahme der Hypothese 3 auch in der Mitte der Skala nur auf wenige Punkte zu.

Die Tabelle 4.6 zeigt diese Daten auszugsweise (vollständig im Anhang Tabelle A.6) und differenziert dazu die Personenparameter entsprechend ihrer Lage in drei Gruppen:

- *Innere Personenparameter* liegen in einem Bereich von jeweils einer halben Standardabweichung über und unter dem Mittelwert der möglichen Personenparameter.

- Personenparameter, die im Bereich von einer halben bis einer Standardabweichung etwas weiter entfernt vom Mittelwert der möglichen Personenparameter liegen, werden als *mittlere Personenparameter* bezeichnet.

- Die außerhalb einer Standardabweichung vom Mittelwert entfernt liegenden Personenparameter werden hingegen als *äußere Personenparameter* bezeichnet.

Für das Beispiel des Mathematik-Testhefts A aus dem Jahr 2015 finden sich mit 48 Items 49 mögliche Personenparametern, von denen die zwei extremen (kein Item und alle Items korrekt gelöst) für diese Untersuchung ausgeschlossen werden. Von den übrigen 47 möglichen Personenparameter fallen 19 in den Bereich der inneren Personenparameter. Alle oben erwähnten 4 Personenparameter, bei denen der schwarze Balken die Nulllinie schneidet, sind genau solche. Für die Bereiche der mittleren und äußeren Personenparameter findet sich im Testheft A des Jahres 2015 kein einziger Personenparameter, für die das Guttman-Pattern in den durch die Hypothese benannten Bereich fällt.

Im Mittel finden sich über alle 34 Itemsets nur 13% der inneren Personenparameter, die zwischen den zwei benachbarten Items liegen, aber in keinem Fall mehr als 36%. Bei den mittleren und äußeren Personenparametern geschieht dies dann mit jeweils etwa 1% noch deutlich seltener. Nur in diesen seltenen Fällen fällt der Personenparameter einer Person mit dem Antwortmuster eines Gutmann-Pattern genau zwischen die zwei Items, welche den Übergang von den gelösten zu den nicht gelösten Items beschreiben.

Die Hypothese 3 muss verworfen werden. Die Rasch-Metrik ist gegenüber Guttman-Pattern nicht unverzerrt.

4.4.4. Zusammenhang der Verteilung von Itemschwierigkeiten und Personenfähigkeiten

Hypothese 4: Durch eine fokussierte Auswahl von Items in einem Schwierigkeitsintervall kann die Verteilung des Messfehlers eines Instruments zielgerichtet beeinflusst werden, so dass es in diesem Intervall besonders sicher misst, folglich dort besonders gut differenziert.

Im Prozess der Entwicklung der VERA-Testhefte wechseln sich Arbeitsschritte ab, die einerseits fachdidaktischen und andererseits psychometrischen Entscheidungen unterliegen (vergleiche auch Kapitel 3). Die Zusammenstellung von Testheften bzw. seit 2020 von Testmodulen ist davon gekennzeichnet, bestimmte Profile von Itemschwierigkeiten herzustellen, mit denen eine Passung zu angenommenen Fähigkeitsverteilungen von Schulformen, Schulen oder

Tabelle 4.6.: Anteil der möglichen Personenparameter, die zwischen den zugeordneten Itemparametern liegen, im Bereich ± einer halben und einer Standardabweichung, sowie dem äußeren Bereich der Verteilung der Personenparameter (Auszug).

				$\pm\frac{1}{2}$ SD			$\pm\frac{1}{2}$ bis 1 SD			außerhalb	
Jahr	Testheft	Items[a]	von[b]	abs.[c]	rel.	von[b]	abs.[c]	rel.	von[b]	abs.[c]	rel.
2008	a	57	22	3	14	18	0	–	16	0	–
2008	b	56	22	8	36	17	0	–	16	0	–
2008	c	42	16	5	31	13	1	8	12	0	–
2009	a	45	18	3	17	14	0	–	12	0	–
2009	b	48	20	2	10	14	0	–	13	0	–
2009	c	40	16	0	–	13	0	–	10	0	–
2010	a	33	13	2	15	11	1	9	8	0	–
2010	b	36	15	1	7	12	0	–	8	1	13
...											
2015	a	48	19	4	21	15	0	–	13	0	–
...											
2019	c	52	21	1	5	16	0	–	14	0	–
2020	b	46	19	4	21	14	0	–	12	0	–
2020	c	39	16	3	19	12	0	–	10	0	–
Mittelwert		1428	571	77	13%	444	4	1%	379	4	1%

Diese Tabelle ist vollständig im Anhang als Tabelle A.6 abgebildet.
[a]Anzahl der Items. Die Zahl der möglichen Personenparameter ist um einen größer. Weil hier aber die zwei äußeren (approximierten) außen vor bleiben, werden folgend nur Anzahl Items minus 1 Personenparameter einbezogen.
[b]Absolute und relative Anzahl der Personenparameter im entsprechenden Bereich.
[c]Anzahl der Personenparameter, die zwischen den zugeordneten Itemparametern liegen.

Klassen hergestellt werden kann. Im folgenden soll dargestellt werden, wie die Verteilungen von Itemschwierigkeiten mit den Verteilungen von Personenparametern zusammenhängen.

Für sämtliche 34 Aufgabensets wurden die Positionen der Aufgaben bezüglich ihrer Schwierigkeit und die Positionen der möglichen Personenparameter in einer Übersicht gegenübergestellt. In der Abbildung 4.12 sieht man die Verteilungen der Item- (blau) und Personenparameter (rot) für die Mathematik-Tests der Jahre 2008 bis 2020 für die jeweils vorliegenden Testheftversionen, bezogen auf die BiSta-Skala. Die Items wurden hier zum besseren Vergleich mit ihrem Parameter bei einer Wahrscheinlichkeit von .50 abgetragen. In der übermittelten Dokumentation des IQB sind die Schwierigkeiten für eine Lösungshäufigkeit von 62,5% bestimmt worden.

Im empirischen Vergleich der Verteilungen lässt sich dabei feststellen:

- Die Mittelwerte der n Itemparameter und der n + 1 möglichen Personenparameter sind identisch[10].

[10]Dies gilt so genau dann, wenn die Schwierigkeit eines Items als der Punkt bestimmt wird, bei dem die

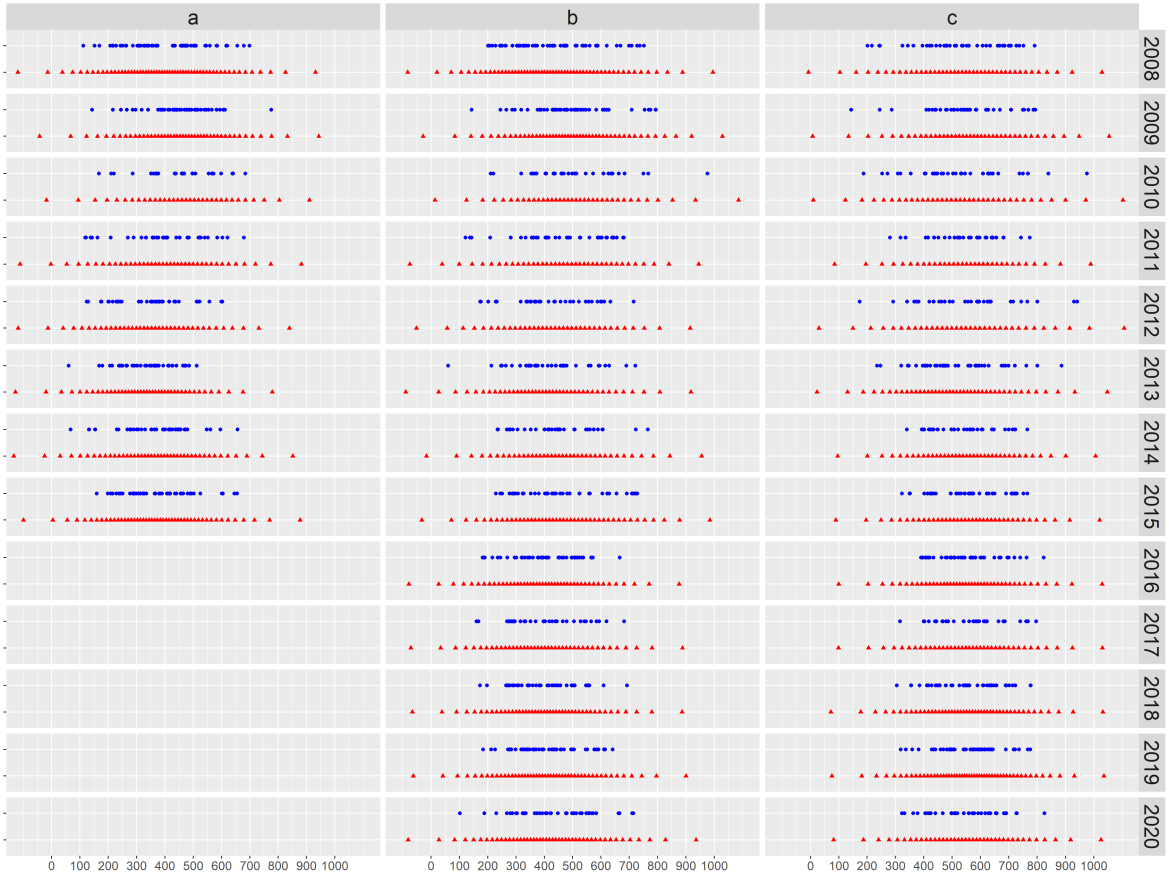


Abbildung 4.12.: Gegenüberstellung der Verteilung von Itemparametern (blau) und Personenparametern (rot) für alle 34 Testhefte von 2008 bis 2020.

- Die Standardabweichung der Personenparameter ist deutlich größer als die der Itemparameter. Eine Regression für die vorliegenden Testhefte zeigt einen mit r = 0.998 fast perfekten Zusammenhang der Form $sd(Personenparameter) = 121,80 + 0,68 * sd(Itemparameter)$.

- Auch Schiefe und Kurtosis der Personenparameterverteilung sind deutlich dichter an einer Normalverteilung, als dies für die Verteilung der Itemparameter gilt.

Fazit: Die Streuung der möglichen Personenparameter ist deutlich breiter als jene der Itemparameter und sie ist bei identischem Mittelwert „normalisiert“. Während die Itemparameter eine Spannweite von im Mittel ca. 550 Punkten bei einer Standardabweichung von fast 100 aufweisen, ist es bei den möglichen Personenparametern eine Spannweite von fast 1.000 Punkten bei einer Standardabweichung von wenig mehr als 50. So sieht man auch in den einzelnen Graphiken der Abbildung 4.12, dass sich die Personenparameter sehr viel ähnlicher verteilen,

Lösungshäufigkeit bei 0,5 liegt. Bei praktischen Realisierungen wird dieser Punkt oft an anderer Stelle positioniert, so für VERA bei 0,625. Damit liegen die zwei Mittelwerte zwar nicht an der gleichen Position, aber in einem festen Abstand zueinander.

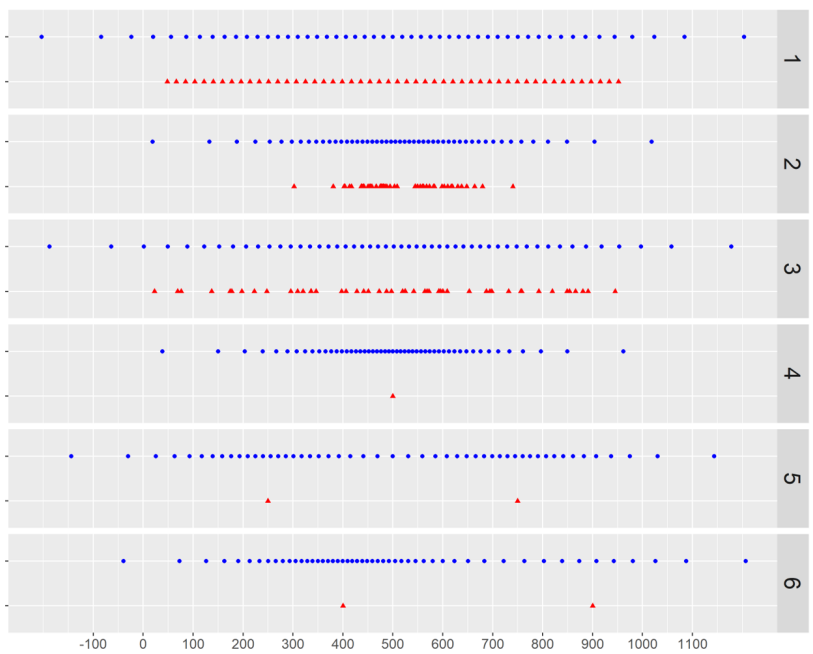


Abbildung 4.13.: Gegenüberstellung der Verteilung von Itemparametern (rot) und Personenparametern (blau) für 6 unterschiedliche simulierte Verteilungen von Itemparametern.

als die Itemparameter. Einzig der Einfluss der zentralen Tendenz ist deutlich. Das C-Testheft aus dem Jahr 2010 weist mit 1093 Punkten die größte Spannweite der Personenparameter auf, was sicher auf die ebenso größte Spannweite der Itemparameter aus diesem Jahr zurückgeführt werden kann. Die kleinste Spannweite nur ein Jahr danach koinzidiert dann auch mit einer der kleinsten Spannweiten der Itemparameter. Einzelne Items können hierbei entscheidend sein, wenn sie wie 2010 die Spannweite durch extreme Randlagen deutlich vergrößern. Die Lagen einzelner Items in der Mitte der Verteilung, selbst einzelne Schwerpunkte scheinen hingegen nur wenig Unterschiede in der Verteilung der Personenparameter zu bewirken.

Für eine nähere Betrachtung dazu, wie sich die Verteilung der Personen- und Itemparameter zueinander verhalten, wurden sechs unterschiedliche Verteilungen von jeweils 50 beispielhaften Items simuliert und die sich ergebende Verteilung der möglichen Personenparameter danebengestellt (siehe Abbildung 4.13).

Sämtliche dieser simulierten Verteilungen von Itemparametern weisen einen Mittelwert von 500 auf, der Mittelwert der Personenparameter ist entsprechend identisch. Man erkennt wieder die deutlich größere Streuung der Personenparameter gegenüber jener der Itemparameter. Selbst bei einer Streuung der Itemparameter von 0 (Verteilung in Beispiel 4), ergibt sich für die Personenparameter eine Streuung von 161. Es zeigt sich, dass die Lage einzelner Itemparameter die Lage der Personenparameter nur marginal beeinflusst. Um tatsächlich eine Häufung

von Personenparameter in einem definierten Bereich der Skala zu erhalten, muss ein relevanter Anteil der Items entsprechend liegen (siehe die Verteilungen der Beispiele 4, 5 und 6). Im Beispiel 5 zeichnet sich erst durch eine extreme zweigipflige Verteilung der Itemparameter eine leichte Häufung um die betroffenen Parameter herum ab. Vergleicht man die Verteilungen der Personenparameter für die Gleichverteilung aller 50 Items aus dem Beispiel 1 einerseits mit der Normalverteilung der Items mit einer Streuung von 250, finden sich nur marginale Unterschiede, wenngleich man in der Normalverteilung durchaus einige Cluster von Items erkennen kann. Ebenso sind die zwei Verteilungen der Personenparameter für die Beispiele 2 und 4 kaum zu unterscheiden, die Verteilung der Itemparameter allerdings deutlich.

In die Berechnung der Lage jedes möglichen Personenparameters gehen sämtliche Itemparameter ein. Allerdings verringert sich der Einfluss der Items deutlich mit der Entfernung zum Personenparameter. Durch die Itemparameter direkt beeinflusst wird zum einen der Mittelwert aller Items, der dem Mittelwert der möglichen Personenparameter identisch ist. Zum anderen können Items in extremen Randlagen die Streuung der Itemparameter erhöhen, was wiederum auch die Streuung der Personenparameter verändert. Die Position der Personenparameter ist allerdings gegenüber den Veränderungen einzelner Itemschwierigkeiten sehr robust.

Die Hypothese 4, nach der sich Häufungen von Itemparametern an bestimmten Positionen der Skala in entsprechenden Häufungen von möglichen Personenparametern widerspiegeln und die Messgenauigkeit genau dort verbessert, muss demnach verworfen werden.

4.4.5. Zusammenfassung und Schlussfolgerungen

Alle hier für die Testinstrumente des Faches Mathematik dargestellten Analysen ergeben sich äquivalent für die Testinstrumente der Domänen Deutsch Lesen und Englisch Leseverstehen, für die für das Land Berlin ebenso lückenlose Ergebnisdaten vorliegen. Deshalb wurde hier auf eine umfassende Darstellung dieser Ergebnisse verzichtet. Nach einer Zusammenfassung der Ergebnisse schließt ein letzter Abschnitt dieses Kapitel mit einem Resümee zur technischen Umsetzung der Simulationen ab.

Zu den Gewissheiten der Rasch-Skalierung

Die vorliegende Auseinandersetzung mit Item- und Personenparametern und ihren Positionen zueinander eröffnet andere als die gewöhnlichen Einblicke in das Funktionieren der Rasch-Skalierung, die eine Schlussfolgerung im Validitätsargument für die Vergleichsarbeiten dar-

stellt (vergleiche Abbildung 2.1).

Durch die Vergleichsarbeiten werden Ergebnisse aus diesen Skalierungen einer breiten Öffentlichkeit offeriert. Lehrkräfte sind aufgefordert aus diesen Ergebnissen Schlussfolgerungen für Ihre Arbeit abzuleiten. Deshalb ist unbedingt nachzuweisen, dass die Daten selbst, aber auch die argumentative Einbettung, so sie für die Interpretation grundlegend ist, zutreffend sind. Aus validitätstheoretischer Perspektive wird die Schlussfolgerung vom beobachteten Antwortvektor auf den beobachteten Fähigkeitswert durch die Rasch-Skalierung, also der Definition der ICCs, begründet, die sich auf bestimmte Voraussetzungen stützt. Im Rahmen der Überprüfung der Hypothese 1 konnte gezeigt werden, dass Items, deren Trennschärfe deutlich von 1 abweichen, insbesondere dann ein Problem für Schlussfolgerungen darstellen, wenn die Abweichungen der Trennschärfen für die Mehrzahl der Items eines Tests in eine Richtung tendieren, wenn also der Mittelwert der Trennschärfen von 1 verschieden ist. Die Größe der Verzerrungen hängt aber nicht nur vom Mittelwert der Trennschärfen ab, sondern auch von der Güte der Passung der Testheftschwierigkeit zur Fähigkeit der Getesteten. Bei guter Passung ist der Einfluss größerer Abweichungen der Trennschärfe weniger oder nicht relevant. Auch wenn die Verzerrungen der Fähigkeitsparameter wegen der von 1 abweichenden Diskrimination in der Realität tatsächlich geringer sind als im simulierten Fall (siehe Tabelle 4.3), sind diese Abweichungen nicht unerheblich. Die Annahme des IQB (siehe Seite 88), dass stärker diskriminierende Items akzeptabler sind als weniger stark diskriminierende scheint hilfreich. Die Auswirkungen auf die Verzerrungen sind aber ähnlich. Deshalb sollte bei der Auswahl von Items Rasch-Konformität im Vordergrund stehen. Die Reduktion dieser Verzerrungen ist zudem ein zusätzliches Argument für die Herstellung einer bestmöglichen Passung zwischen Testheftschwierigkeit und Personenfähigkeit.

Im Abschnitt 4.1.2 der Hypothesenprüfung 2 erkennt man in der Darstellung 4.9 sehr gut, mit welch großem Konfidenzintervall individuelle Personenparameter belegt sind. In den Rückmeldungen, die zum Beispiel durch das ISQ für die Schulen in Berlin und Brandenburg bereitgestellt werden, wird dieses Konfidenzintervall mit einem Grauverlauf kommuniziert (siehe Abbildung 4.14).

Dargestellt sind jeweils am linken und rechten Rand die individuellen Kompetenzstände durch einen weißen Strich in einem grauen Band. Die weiße Linie ist an der nebenstehenden BiSta-Skala ausgerichtet, die durch die Teilung in die Kompetenzstufen der Domäne abgebildet wird. Die Breite aller inneren Stufen ist identisch. Für ein angenehmes Bild sind die jeweils nach unten bzw. oben offenen Stufen an den Rändern mit der gleichen Breite dargestellt. Im

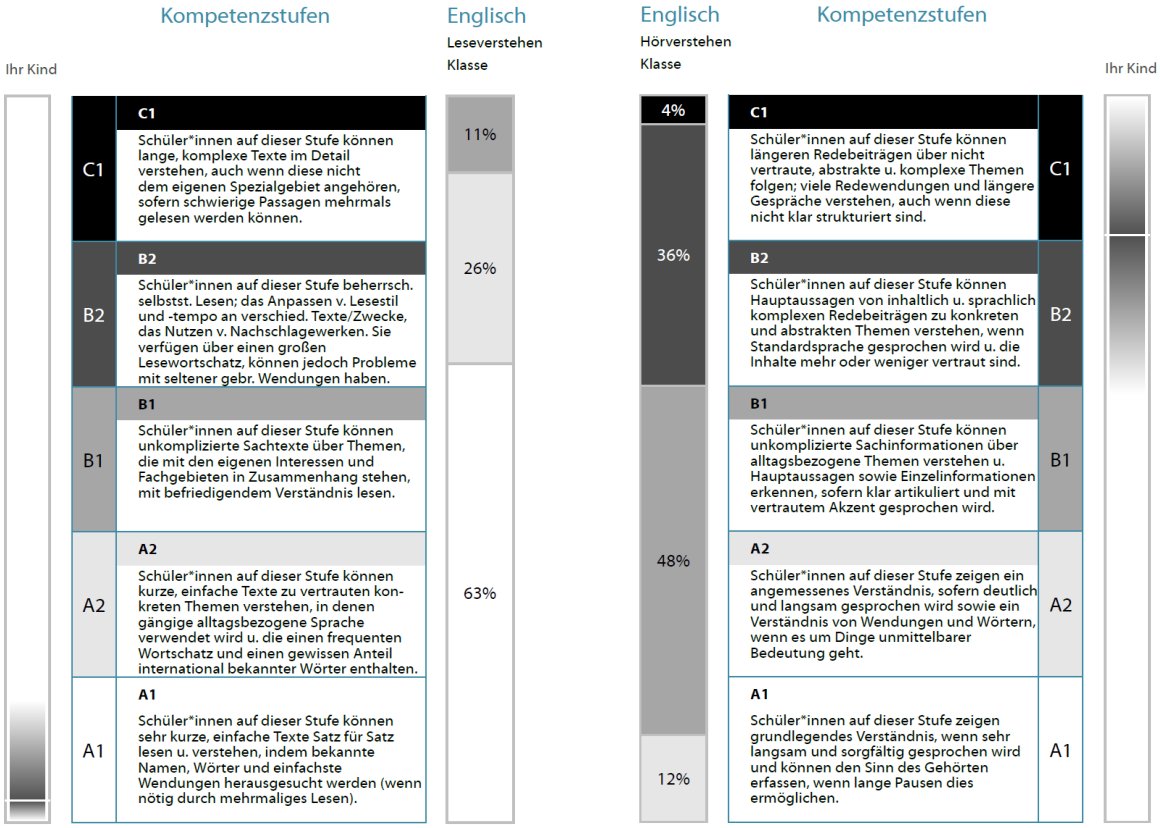


Abbildung 4.14.: Individuelle Kompetenzstände im kriterialen und sozialen Vergleich in einer Rückmeldung für das Fach Englisch bei VERA-8, 2020 im ISQ. In der Rückmeldung stehen die zwei Rückmeldungen für die Domänen Leseverstehen (links) und Hörverstehen (rechts), wie hier abgebildet, nebeneinander. Wegen der Veranschaulichung verschiedener Effekte ist hier aber eine Graphik für das Leseverständnis aus einem Test mit dem einfacheren A-Heft aber eine Hörverstehens-Rückmeldung aus einem Testheft B dargestellt.

linken Beispiel der Domäne Englisch Lesen liegen die Grenzen zwischen den Stufen bei 400, 500, 600 und 700 BiSta-Punkten. Erst mit 8 von 38 richtig gelösten Aufgaben erreicht man 312 BiSta-Punkte und damit den durch die Graphik abgebildeten Skalenbereich von 300 bis 800. Geringere Lösungshäufigkeiten werden sämtlich am unteren Rand dargestellt (siehe das Beispiel für Englisch Leseverstehen in der Abbildung 4.14 links). Am oberen Rand betrifft dies nur einen Wert bei 865 Punkten, der als Randwert dargestellt wird. Hier abgebildet ist die einfache Version des Testhefts aus dem Jahr 2020, welches vermehrt Aufgaben im unteren Kompetenzbereich enthält. Für das schwierigere Testheft wird im unteren Bereich nur der Fall von einem von 31 korrekt gelösten Aufgaben künstlich in die Darstellung „verschoben“, während es im oberen Bereich zwei mögliche Personenparameter betrifft. Diese Asymmetrie rührt zu einem Teil natürlich daher, dass die Kompetenzstufen für das Ende der Klassenstufe 10 entwickelt wurden. Zu diesem Zeitpunkt sollte sich das Gros der Schüler*innen mit dem

Regelstandard in der Mitte der dargestellten Kompetenzstufen befinden.

Die weiße Linie für den individuellen auf die BiSta-Skala transformierten WLE-Parameter der Schülerin bzw. des Schülers steht in der Mitte eines Grau-Verlaufs, welcher das Konfidenzintervall abbildet. Wegen des oben beschriebenen Effektes ist der links zurückgemeldete Personenparameter tatsächlich eher aus der Mitte einer Messung, während der rechte Parameter aus einem Randbereich der Messung stammt. Daraus resultieren die deutlich verschieden breiten Konfidenzintervalle. Am Rand einer Rasch-skalierten Messung sind diese deutlich größer. Der Rand des Konfidenzintervalls ist hier allerdings nicht deutlich zu erkennen. Zum einen hätte ein exakter Rand eine Diskussion um die Genauigkeit der Messung geradezu provoziert, die zu einem Großteil ohne entsprechendes Hintergrundwissen zur Berechnung von Konfidenzintervallen geführt worden wäre. Andererseits wird bei Harych und Emmrich (2019) die Frage diskutiert, inwieweit ein 95%-Konfidenzintervall für den schulpraktischen Kontext, in dem die Ergebnisse interpretiert werden, eine sinnvolle Wahl darstellt. Deshalb wurde als Kompromiss ein Grau-Verlauf gewählt, dessen Nichtlinearität unterschiedlich breite Konfidenzintervalle berücksichtigt[11].

Trotz des großen Konfidenzintervalls scheint es unangemessen, individuelle Ergebnisse von einer Interpretation grundsätzlich auszuschließen. Es ist davon auszugehen, dass viele andere individuelle Lernstandsanalysen mit ähnlichen Unsicherheiten behaftet sind, letztendlich Lehrkräfte aber damit umzugehen wissen. Einzelne Ergebnisse aus Lernstandsanalysen wie auch aus Vergleichsarbeiten sind sinnvoll mit anderem Wissen um die Leistungsfähigkeit der Schüler*innen zu verknüpfen und dabei deren Kontextabhängigkeit und Unschärfe bewertend einzubeziehen. Allerdings machen Harych und Emmrich (2019, S.281) deutlich, dass der Umgang mit Unsicherheit nicht nur wenig problematisiert, sondern geradezu ausgeblendet wird und deshalb für Lehrkräfte herausfordernd ist. Die Autoren fragen sich als Teil der Testadministration aber auch selbst, inwieweit die unreflektierte Übernahme eines 95%-Konfidenzintervalls für den Kontext der Unterrichtsentwicklung angemessen ist (ebenda, S.279).

Dass das Guttman-Pattern nur selten jenen Platz einnimmt, der ihm beim Standard-Setting eingeräumt wird, ist misslich. Insbesondere größere Abweichungen, wie sie die Prüfung der Hypothesen 3 offeriert, sind hier schwer zu erklären und befördern Misstrauen. Vielleicht ist zu prüfen, ob eine Metrik mindestens der Anforderung genügen kann, die für Guttman-

[11] Der gesamte Balken des Grau-Verlaufs bildet das 95%-Konfidenzintervall ab, wobei die Enden durch den Verlauf in die Hintergrundfarbe Weiß nicht präzise auszumachen ist. Zudem verläuft das Grau über die Skala nicht linear ins Weiß, sondern zuerst schneller. Damit wird der Nichtlinearität entsprochen, die sich darin ausdrückt, dass ein 45%-Konfidenzintervall kleiner ist als ein halbes 90%-Konfidenzintervall.

ähnliche Pattern Unverzerrtheit verlangt. Solche Pattern zeichnen sich dadurch aus, dass in einer nach der Schwierigkeit geordneten Menge an Items vom ersten Item I_1 bis zum Item I_n alle richtig gelöst werden und ab einem bestimmten Item I_{n+m} sämtliche folgenden bis zum letzten Item I_z nicht korrekt gelöst werden konnten. Der Personenparameter sollte dann zwischen den beiden Items I_n und I_{n+m} liegen. Das Guttman-Pattern wäre dann der Spezialfall mit $m = 1$. Mindestens für Werte von m ab einer gewissen Größe sollte die Forderung erfüllbar sein. Andernfalls wäre es bei den ohnehin vorliegenden großen Fallzahlen auch möglich, statt des Rasch-Modells ein 2PL-Modell zu verwenden. Die beschriebenen schwer erklärbaren Probleme würden sich damit offenbar erübrigen. Dieser Vorteil wäre allerdings gegen die Nachteile abzuwägen. Wie schon im Abschnitt 3.3 erwähnt, nutzt die OECD für PISA seit 2015 auch 2PL-modellierte Items. Ein Guttman-Pattern als reales Antwortmuster und die dazu unpassende Position der beobachteten Fähigkeit einer Schülerin oder eines Schülers werden unterrichtspraktisch wohl kaum identifiziert und hinterfragt. Die offenbarte Differenz widerspricht eher der Argumentation für das Setzen von Cut-Scores beim Standard-Setting. Vergegenwärtigt man sich allerdings, wie groß der Anteil an Fehlklassifikationen bei der Zuordnung von Schüler*innen zu Kompetenzstufen ohnehin ist (Pant et al., 2017, S.63), sind die hier gezeigten Differenzen relativ klein, so dass die Einschränkung der Validität hinnehmbar erscheint.

Die Prüfung der vierten Hypothese verdeutlicht, dass die tatsächliche Verteilung von Itemschwierigkeiten für ein Testheft wenig Einfluss auf die Position möglicher Personenparameter hat. Lediglich die mittlere Schwierigkeit verweist auf den Bereich der Fähigkeitsskala, für den die Messgenauigkeit optimiert ist. Eine Testheftauswahl auf die mittlere Lösungshäufigkeit in einer Population zu stützen, scheint demnach ein sinnvolles Vorgehen zu sein. Darüber hinaus scheint es für die Auswahl eines Testhefts eher angezeigt, die konkrete inhaltliche Repräsentation mit dem realisierten Curriculum abzugleichen.

Die vorliegenden Analysen stützen sich ausschließlich auf Simulationen, die wiederum lediglich die 34 aus den vergangenen Jahren vorliegenden Sets von Items nutzen. Es konnten einige Einschränkungen bei der Interpretation von Ergebnissen aus WLE-Parameterschätzungen herausgestellt werden, welche die ermittelten Fähigkeitszuschreibungen mit einigen (zusätzlichen) Unsicherheiten belegen. Für Fähigkeitsschätzungen von Populationen sind plausible values als Schätzer zu favorisieren. Im Kontext von Lernstandsfeststellungen ist eine Zuschreibung von Fähigkeiten für einzelne Schüler*innen und damit die Nutzung von WLE-Schätzungen aber nicht zu umgehen. Die Verwendung von auf dieser Basis aggregierten Er-

gebnissen führt aber ggf. zur Fortschreibung der hier dargestellten Problemlagen bei der Interpretation. Allerdings bleibt unklar, ob diese Probleme für eine Nutzung der WLE-Ergebnisse im schulpraktischen Kontext von Bedeutung sind. Dies wäre weiter zu untersuchen.

Zur technischen Umsetzung der Replikationen

Weil noch im Jahr 2016 Feinberg und Rubright beklagten, dass in Artikeln, deren Ausführungen sich auf Simulationen stützen, zumeist keine oder nur sehr wenige Angaben zur technischen Umsetzung der Simulationen zu finden seien, wurden in der vorliegenden Arbeit solche Informationen in den entsprechenden Abschnitten eingeflochten. Zusammenfassend kann hier festgestellt werden, dass sich mit der modernen Syntax von R (R Core Team, 2021) Simulationen sehr gut umsetzen lassen. Mit der auf Servern verfügbaren Webapplikation von RStudio können zudem die skalierbaren Ressourcen von virtuellen Servern genutzt werden, was sich gleichermaßen auf die Zahl der CPUs bzw. Rechenkerne bezieht wie auch auf den Arbeitsspeicher. Überdies ist hier ein dauerhafter Betrieb am einfachsten zu gewährleisten. Auch wenn die weitgehend Rasch-konformen Datensätze keine besonderen Herausforderungen darstellten, konkret keine Konvergenzprobleme erwarten ließen, denen mit vielen Iterationen hätte begegnet werden müssen, überzeugt die Geschwindigkeit von R mit über einer halben Million Rasch-Skalierungen innerhalb von 35 Stunden (mit nur einem Kern und 16GB RAM). Diese Performance zeigte sich äquivalent auch für die Simulationen der zwei weiteren Domänen. Grundlage dafür ist sicher auch die hohe Performance und Stabilität des eingesetzten R-Moduls TAM (Robitzsch et al., 2020) für sämtliche Rasch-Skalierungen.

Für die programmatische Umsetzung großer und größerer Projekte sollte von Beginn an auf Lösungen gesetzt werden, die in erster Linie den Arbeitsspeicher so wenig wie möglich belasten. Mit der seed-Funktion können wiederholbare Zufallsauswahlen erzeugt und repliziert werden, und so auch zufällige Daten besser wiederholt produziert, als permanent im Arbeitsspeicher gehalten werden. Dann lohnt sich auch, auf die parallele Bearbeitung mit mehr als einem Kern zu setzen und dazu die entsprechenden Bibliotheken einzubinden. Zudem eröffnet eine vollständige Replizierbarkeit selbst (pseudo-)zufälliger Prozesse eine einfache Möglichkeit der Kontrolle durch Dritte.

5. Stabilität der Ergebnisse von Vergleichsarbeiten

Die in diesem Kapitel vorgestellte Studie untersucht die Validität der Interpretationen der Ergebnissen von Vergleichsarbeiten aus unterschiedlichen Jahren. Wegen des Bezugs aller Ergebnisse auf die domänenspezifisch gemeinsame Skala der Bildungsstandards, sind solche Schlussfolgerungen naheliegend. Mit der Gültigkeit dieser Interpretationen wird jener Teil des Validitätsargumentes überprüft, mit dem von den beobachteten auf die erwarteten Ergebnisse geschlossen wird. Durch das Linking sind die verschiedenen Testhefte als Instrumente für die Kompetenzfeststellung des identischen Konstrukts anzusehen. Eine gemeinsame Interpretation stellt somit eine Generalisierung über verschiedene Testkontexte dar.

Trendbezogene Interpretationen von Ergebnissen aus Vergleichsarbeiten präsentiert der erste Abschnitt 5.1 und spiegelt damit exemplarisch solche Diskussionen. Im zweiten Abschnitt 5.2 werden mit Panel- und Trendstudien zwei Studienarten für Veränderungsmessungen unterschieden, bevor im Abschnitt 5.3 der Frage nachgegangen wird, wie viel Stabilität denn bei den Vergleichsarbeiten überhaupt erwartet werden kann. Die zwei zentralen Abschnitte nähern sich dem Gegenstand der Validität auf zwei unterschiedlichen Wegen. Erweisen sich im Abschnitt 5.4 untersuchte Artefakte im Rahmen von VERA als unplausibel, beeinträchtigt dies die Feststellung von Validität. Die Ergebnisse werden dabei teilweise mit Rückgriff auf berichtete Messungen beim Bildungstrend abgeglichen. Im Weiteren wird mit Daten aus zwei ISQ-Studien im Abschnitt 5.5 untersucht, inwieweit der wiederholte Einsatz von VERA-Instrumenten erwartbare Ergebnisse zeitigt bzw. unter welchen Bedingungen sich zuvor gezeigte Artefakte wiederholen. Abgeschlossen wird dieses Kapitel mit zusammenfassenden Schlussfolgerungen (5.6).

5.1. Interpretationen von Trends

Ab dem Schuljahr 2016/17 wurde über die Ergebnisse der Vergleichsarbeiten für die Länder Berlin und Brandenburg nicht mehr öffentlich Bericht erstattet. Darauf einigte man sich nach Empfehlung des Wissenschaftlichen Beirats des ISQ mit den Administrationen beider Länder. Vorher waren insbesondere die Berliner Ergebnisse jährlich Thema in Artikeln der Presse bzw. von Veröffentlichungen der Presseabteilung der Senatsbildungsverwaltung. Im Allgemeinen wurde hier das schlechte Abschneiden beklagt und dort Verbesserungen hervorgehoben, in vielen Fällen als Veränderungen zum Vorjahr. 2020 fanden sich äquivalente Interpretationen als Trends im Abschlussbericht der von der Senatsverwaltung einberufenen Expertenkommission (Köller et al., 2020). Köller et al. (2020, S.87) präsentieren hier die Anteile von Schülerinnen und Schülern, die bei VERA-8 lediglich die Kompetenzstufe I in den Domänen Deutsch-Lesen und Mathematik erreicht haben (identische Daten, wiedergegeben mit eigenen Abbildungen 5.1 und 5.2), und sie interpretieren schulformspezifische Trends

- für Deutsch Lesen als „deutliche“ Zunahme für die Gruppe der nicht-gymnasialen Schulformen (Integrierte Sekundar- und Gemeinschaftsschulen),
- für die Gymnasien hingegen als „weitgehend konstant niedrig“ und
- für Mathematik als „negative[n] Trend“, der „in beiden Schulformen leicht ansteigend“ ist.

Zuzustimmen ist der zusammenfassenden Darstellung der Autor*innen, nach der „vor allem im nicht-gymnasialen Bereich große Anteile der Schülerinnen und Schüler im Lesen und in Mathematik den Vorgaben der KMK weit hinterherhinken“ (ebenda, S. 87), womit aber lediglich auf einen, vermutlich den mit VERA-8 2019 aktuell ermittelten, Leistungsstand Bezug genommen wird. Mindestens in Anbetracht der Untersuchungen aus dem vorigen Kapitel ist aber die Frage zu stellen, ob Interpretationen von Trends für Zeitreihen aus Messungen des Leistungsstands mit VERA als valide gelten können. Für die Nutzung der Anteile von Kompetenzstufen haben die vorgängigen Untersuchungen in dieser Arbeit ausreichend Zweifel hinterlassen. Selbst in Anbetracht der als Vollerhebung zu wertenden Untersuchung, für Mathematik tatsächlich von einem leicht negativem Trend zu sprechen, scheint da eher unangemessen. Die Zunahme des Anteils von Schülerinnen und Schülern nicht-gymnasialer Schulformen auf Kompetenzstufe I für die Domäne Deutsch Lesen innerhalb der Jahre von 2014 bis 2018 von 21% auf 52% scheint allerdings ein dramatischer Befund zu sein. Wie der

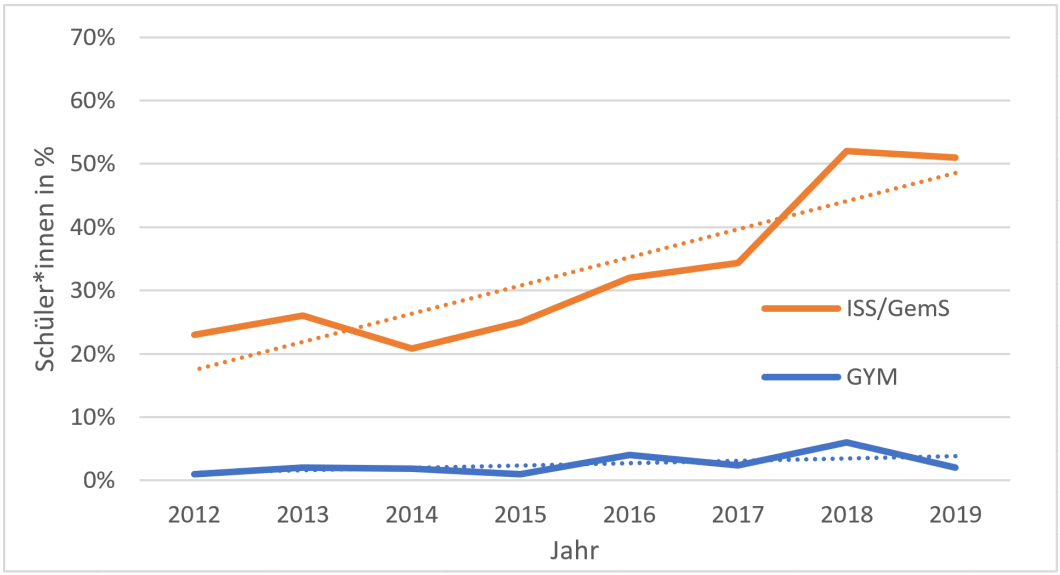


Abbildung 5.1.: Leistungsschwache Leserinnen und Leser (Kompetenzstufe I) in VERA 8 nach Schulform und Erhebungsjahr

Bildungstrend liefern solche Darstellungen allerdings keinerlei Erklärungen für solche Befunde.

Noch dramatischer als der Befund selbst ist aus schulpraktischer wie bildungspolitischer Sicht, dass eine derartige Steigerung des Anteils von Schülerinnen und Schülern mit Leistungen, die nur der Kompetenzstufe I zugerechnet werden, quasi unbemerkt bleibt. Ein Anstieg des Anteils auf das Zweieinhalbfache innerhalb von nur vier Jahren bei Schüler*innen, für die nach Köller et al. (2020) „die Sorge besteht, dass diese Jugendlichen [...] Probleme haben werden, anschlussfähig zu lernen“ (S.87). Ist eine solche Veränderung ein tatsächlich erwartbarer, ein erklärbarer Befund? Und wenn solche unterschiedlichen Leistungsstände reale Entwicklungen abbilden, welche Konsequenzen hat dies für die im fünfjährigen Turnus stattfindende Messung des Bildungstrends?

Unabhängig von der hier zitierten Ergebnisdarstellung stellen aber auch andere Akteure Bezüge zwischen VERA-Rückmeldungen verschiedener Jahre bzw. zwischen diesen und anderen als weitgehend äquivalent angesehenen Messungen her. Alle VERA-Instrumente werden mit Bezug zu den Bildungsstandards entwickelt. Was bedeutet dies konkret? In der ersten Gesamtstrategie der KMK (KMK, 2006b) wurde bei den Ausführungen zum Ländervergleich und späterem Bildungstrend durch das IQB eine *Anbindung* an die internationalen Erhebungen mit Hilfe von Ankeraufgaben avisiert (ebenda S.22). Ziel war über die Nutzung identischer oder zumindest äquivalenter inhaltlicher Konzepte hinaus auch eine psychometrische Verknüpfung die sicherstellt, dass die Ergebnisse quantitativ aufeinander bezogen werden

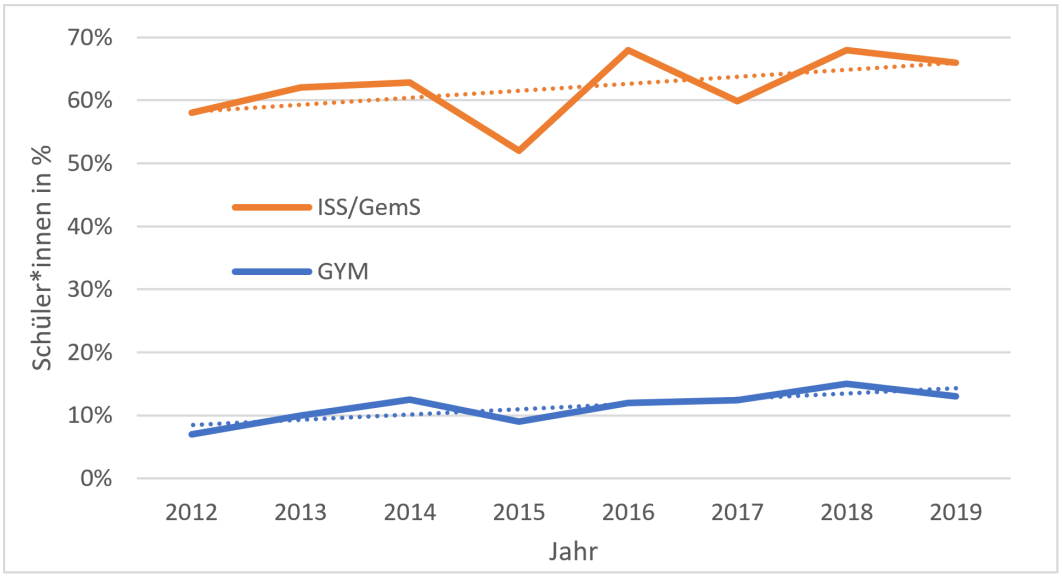


Abbildung 5.2.: Schwache Schülerinnen und Schüler in Mathematik (Kompetenzstufe I) in VERA 8 nach Schulform und Erhebungsjahr

können. Das selbe Dokument verweist für die Vergleichsarbeiten auf den Unterschied zwischen *Anlehnung* und *Ankopplung* und beschreibt dazu Anlehnung als eine nur[1] „inhaltliche Orientierung“. Ankopplung wird darüber hinaus als psychometrische Verknüpfung, äquivalent zur Anbindung, beschrieben. Gleich anschließend wird im Falle einer Ankopplung festgestellt, dass „die Ergebnisse von Vergleichsarbeiten direkt mit den Befunden des zentralen Ländervergleichs zur Überprüfung des Erreichens der Bildungsstandards verknüpft werden“ (ebenda, S.21) können. Dies führt dann natürlich dazu, dass die Ergebnisse verschiedener Vergleichsarbeiten auf einer identischen Metrik verortet und damit in Bezug zueinander gesetzt werden können. Damit wird eine Stabilität über eine einzelne Messung hinaus postuliert, welche Basis der oben festgestellten Interpretationen ist bzw. überhaupt erst eine Interpretation solcher Veränderungen erlaubt.

Die Untersuchung von Stabilität meint die Detektion, welche Anteil einer Veränderung der tatsächlichen Variation des gemessenen Konstrukts zuzuschreiben sind. Mit Bezug zum argumentativen Ansatz von Kane (2013) wird damit die Gültigkeit der Generalisierung (vergleiche Abbildung 2.1) von Ergebnissen über verschiedene Sets von Aufgaben überprüft. Keine der referierten Studien hatte die Untersuchung der Stabilität der Messung der Vergleichsarbeiten als originäres Ziel benannt. Die Wahl der Instrumente für den dritten Messzeitpunkt der VERAMSA-Studie (Graf et al., 2016) kann allerdings als Indiz für diesbezügliche Unsicherheiten gelten. Trotzdem ist unklar, weshalb eine veröffentlichte Betrachtung der zeitlichen

[1] Dieses „nur“ wurde hier ergänzt, weil dies die Intention zutreffend beschreibt.

Stabilität der VERA-Instrumente bislang fehlt, wo doch deutlich wird, dass Auffälligkeiten nicht auf die Messungen in Berlin und Brandenburg beschränkt sind. Dieses Kapitel ist eine Bestandsaufnahme für zukünftige Untersuchungen.

5.2. Nutzung der Ergebnisse aus VERA als Panel- und/oder Trendstudie

Für quantitative Aussagen zur Entwicklung von beispielsweise Schülerleistungen werden im Allgemeinen an mindestens zwei Zeitpunkten, welche den Entwicklungszeitraum angemessen abbilden, entsprechende Messungen durchgeführt[2]. Es lassen sich zwei grundlegende Formen solcher längsschnittlicher Untersuchungen unterscheiden (Reinecke, 2012, S.21).

Für Trendaussagen, also Aussagen zu Veränderungen auf beliebig aggregierter Ebene, werden in einem bestimmten Turnus mit äquivalenten (oder als äquivalent anzusehenden) Instrumenten jeweils unterschiedliche Stichproben einer Grundgesamtheit untersucht. So erhob das IQB zum Beispiel im Rahmen des Bildungstrends 2012 und 2018 die Mathematikkompetenzen von Schülerinnen und Schülern aus neunten Klassen (Pant et al., 2013; Stanat et al., 2019a) und zog dazu in beiden Jahren jeweils eine repräsentative Stichprobe, die notwendig jeweils andere Schülerinnen und Schüler einschloss. Das verwendete Instrumentarium zur Messung der Kompetenz blieb dabei im Wesentlichen identisch. Die Ergebnisse der zweiten Erhebung erlauben einerseits Aussagen über den Leistungsstand der Neuntklässler*innen im Jahr 2018, gleichsam einer einmaligen Querschnittsmessung. Darüber hinaus können aber auch Aussagen zur Entwicklung der Leistung von Neuntklässler*innen in den vergangenen 6 Jahren formuliert werden. Die Aussage zur Entwicklung betrifft gerade nicht die individuelle Entwicklung einzelner Schülerinnen und Schüler und auch nicht die Entwicklung einer abgegrenzten Gruppe von ihnen, sondern die Leistungsstände zweier voneinander disjunkter Gruppen von Schülerinnen und Schülern, die das Merkmal teilen, im jeweiligen Erhebungsjahr die neunte Klasse besucht zu haben. Die Gründe für eine in solch einer *Trendstudie* festgestellten Entwicklung sind demnach außerhalb der Individuen zu suchen, zum Beispiel in einer durch überproportionale Zu- oder Abwanderung veränderten Population, oder auch als Folge von in diesem Zeitraum wirksam gewordener Reformen, vielleicht auch in einem hohen Unterrichtsausfall als Folge pandemiebedingter Schulschließungen. Eine klassische Trendstudie ist natürlich der Bildungstrend, der im ersten Zyklus noch Ländervergleich hieß. Berichte des

[2]Im Folgenden wird immer der Fall von zwei Messzeitpunkten beschrieben, wobei die Erläuterungen in identischer Weise auch für jede größere Zahl von Messungen zutreffen.

zweite Zyklus, die auch Trendaussagen enthalten, liegen für die Sekundarstufe mit Stanat et al. (2016) und Stanat et al. (2019a) und für die Primarstufe mit Stanat et al. (2017a) vor. Wie von den Autoren selbst deklariert, stellen die oben wiedergegebenen Darstellungen aus dem Expertenbericht (Köller et al., 2020) ebenso Trends mit Hilfe von auf Landesebene aggregierten VERA-Ergebnissen dar. Auch, wenn es naheliegend scheint, die jährlichen Vollerhebungen der fachlichen Kompetenzen im Rahmen der Vergleichsarbeiten als Trend darzustellen, finden sich solche Untersuchungen kaum. Im Bildungsbericht Bayerns von 2018 (Lankes et al., 2018, S. 91) wurden für Deutsch-Lesen und Mathematik Ergebnisse der Vergleichsarbeiten der dritten Jahrgangsstufe von zwei Jahren gegenübergestellt. Ergebnisse von Vergleichsarbeiten werden also nur selten in Trenddarstellungen verwendet, gelegentlich aber, wie eingangs gezeigt, als solche interpretiert.

Von solchen Trendstudien sind *Panelstudien* zu unterscheiden. Hier wird eine identische Stichprobe von Schülerinnen und Schülern, also identische Entitäten, zu mindestens zwei Zeitpunkten untersucht. Dies geschieht mit der Zielstellung eine Entwicklungsaussage für genau die untersuchte Kohorte von Schülerinnen und Schülern innerhalb des Untersuchungszeitraumes zu treffen. Systemische Veränderungen werden hier nur dann abgebildet, wenn sie in genau dem Untersuchungszeitraum, genau die Mitglieder*innen der untersuchten Population betreffen. Untersucht werden können mit Panelstudien allerdings individuelle Entwicklungsverläufe. Der Einsatz eines identischen Messinstruments ist aber ggf. mit spezifischen Problemen verbunden. Liegen zum Beispiel bei einer Kompetenzmessung die zwei Messzeitpunkte dicht beieinander, können sich die untersuchten Schüler*innen vielleicht an die Aufgaben erinnern. Bei Messungen über einen großen Zeitraum können die erwarteten Veränderungen wiederum so groß sein, dass zwingend andere Aufgaben verwendet werden müssen. Werden aber nicht die gleichen Instrumente verwendet, müssen geeignete Maßnahmen ergriffen werden, um sicherzustellen, dass in beiden Fällen das gleiche Konstrukt gemessen wird. Im Fall von Kompetenzentwicklungen (über einen größeren Zeitraum) muss dann ggf. ein Kompetenzentwicklungsmodell zu Grunde gelegt werden, dessen Entwicklung andere Herausforderungen mit sich bringt. In der Schweiz werden die zentralen Erhebungen zur Überprüfung der Grundkompetenzen (ÜGK) in einzelnen Kantonen bzw. Verbünden mit solchen, an Kompetenzentwicklungsmodellen orientierten Instrumenten begleitet (siehe www.stellwerk-check.ch und www.check-dein-wissen.ch).

Der grundlegenden Intention der Vergleichsarbeiten folgend, die sich auf die Analyse individueller Leistungsfeststellung bezieht bzw. auf den Leistungsstand einer Klasse, finden sich in

diesem Zusammenhang verschiedene Panelstudien. Eine klassische Variante ist die Untersuchung der Entwicklung der Orthographiekompetenz von Schülerinnen und Schülern zwischen den Klassenstufen 3 und 4 (Vettorazzi & Harych, 2019). Die mit den Vergleichsarbeiten in der dritten Klasse gemessenen Kompetenzstände, wurden dabei ein Jahr später an einem Teil der Population der dann Viertklässler*innen mit dem identischen Test erneut überprüft. Hier kann für jede Schülerin und jeden Schüler ein Leistungszuwachs ermittelt werden. Eine weitere ISQ-Studie VERAMSA (Graf et al., 2016) sollte die Prognosegüte der Vergleichsarbeiten in der achten Klasse für das Ergebnis der Schülerinnen und Schüler beim Mittleren Schulabschluss (MSA) untersuchen. Dazu wurde in der untersuchten Stichprobe nach dem regulären Test im Rahmen der Vergleichsarbeiten der Klassenstufe 8 ein weiterer Test mit Testheften aus den Vergleichsarbeiten in der neunten Jahrgangsstufe und kurz vor dem MSA in der zehnten Jahrgangsstufe durchgeführt. Hier liegt dementsprechend ein Datensatz einer Panelstudie mit drei Messzeitpunkten vor. Die Ergebnisse der Prüfungen zum Mittleren Schulabschluss ergänzen diese Daten, allerdings sind die Prüfungsergebnisse nicht auf der Metrik der Bildungsstandards verortet, wie das allerdings für jedes VERA-Testheft der Fall ist. In Hamburg verbindet KERMIT (Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), 2021; Thonke & Lücken, 2014; Benzig et al., 2016) Lernstandsmessungen aus den Jahrgangsstufe 2, 3, 5, 7 und 9 über eine gemeinsame Metrik miteinander und ist damit vermutlich eine der größten Panelstudien Deutschlands. Für die Tests in den Jahrgangsstufen 3 und 8 werden dabei die Vergleichsarbeiten eingebunden. Für die anderen Jahrgangsstufen kommen in Hamburg selbst entwickelte Instrumente zum Einsatz, die mit der Metrik der Vergleichsarbeiten verlinkt werden. Im Prinzip startet hier mit jedem Jahr ein neues Panel mit der Testung einer zweiten Klasse. Die Vergleichsarbeiten werden also in besonderen Szenarien in Panelstudien eingebunden.

Neben dem Unterschied ob zu den unterschiedlichen Erhebungszeitpunkten in einer Panelstudie die identischen Entitäten untersucht werden oder unterschiedliche Entitäten in einer Trendmessung, ist zu beachten, ob für die Messungen an den verschiedenen Zeitpunkten identische oder unterschiedliche Instrumente verwendet werden. Unterschiedliche Instrumente müssen entweder über eine gemeinsame Metrik verfügen, so dass die Ergebnisse als identisch betrachtet werden können. Ist dies nicht der Fall, muss eine Vorschrift gefunden werden, mit der die zwei unterschiedlichen Metriken ineinander überführt werden können[3]. Durch die gemeinsame Entwicklung der Instrumente und deren Verlinkung durch das IQB werden die

[3]Dieser Vorgang wird auch als Equating bezeichnet.

Ergebnisse von Vergleichsarbeiten und auch jene der Bildungstrend-Untersuchungen auf einer identischen Metrik abgebildet, sind somit also grundsätzlich aufeinander beziehbar. Bei der Untersuchung zum Kompetenzzuwachs in Orthographie von der dritten zur vierten Klasse wird beispielsweise der identische Test eingesetzt, so dass sich die Ergebnisse sehr einfach interpretieren lassen. Wenn bei VERAMSA Ergebnisse von VERA-8 Ergebnissen gegenübergestellt werden sollen, die zwei Jahre später mit einem anderen VERA-8-Testheft ermittelt wurden, dann müssen die zwei Testhefte bzw. die zwei Messungen miteinander verlinkt werden. Da für zwei VERA-8-Instrumente die gemeinsame Metrik der Bildungsstandards verwendet wird, ist diese Verlinkung als gegeben hinzunehmen. Demgegenüber gestaltete sich die Feststellung des Leistungszuwachses zwischen dem mit VERA-8 festgestellten Leistungsstand in der 8. und dem Ergebnis der Prüfung zum Ende der 10. Klasse im Projekt VERAMSA als deutlich schwieriger. Hier musste eine Verlinkung nachträglich hergestellt werden.

5.3. Erwartete Stabilität

Im Bildungstrend des IQB werden die Kompetenzstände von Schülerinnen und Schülern erhoben, die im Jahr der Erhebung für den Bildungstrend der Primarstufe eine vierte und für den Bildungstrend in der Sekundarstufe eine neunte Klasse besuchen. Differenzen zwischen jeweils zwei Erhebungen werden, sofern sie sich als überzufällig erweisen, als Ergebnis von systemischen Veränderungen, wie Bildungsreformen, der Implementation von Kompetenzorientierung, der Implementation eines neuen Qualitätsmanagements durch Schulinspektorate etc., interpretiert oder sie werden auf grundlegende Veränderungen in der Population zurückgeführt. Diese Zuschreibungen finden nur sehr begrenzt im jeweiligen Trendbericht statt, sondern sind Teil der Analyse des Berichts in den einzelnen Bildungsadministrationen. Differenzen die nicht einer Veränderung des wahren Wertes zugeschrieben werden können, werden im Allgemeinen Quellen zugeschrieben, deren Einwirken auf die Messung als *zufällig* beschrieben wird. Das sind zum einen zumindest in ihrer Größe beeinflussbare Quellen wie Stichproben- und Messfehler. Zum Anderen muss aber auch davon ausgegangen werden, dass die Wirklichkeit Stochastizität produziert, deren vollständige Aufklärung nicht möglich ist. Die hiermit verbundene philosophische Diskussion soll hier nicht geführt werden. Es kommt für eine Abschätzung der Stabilität der Messung darauf an, die Veränderungen des wahren Wertes einerseits und die Größe der Stichproben- und Messfehler sowie des „Rauschens“ der Wirklichkeit andererseits in ihren Größenordnungen abzuschätzen.

Dass sich über die Jahre Veränderungen des Kompetenzniveaus bei den Vergleichsarbei-

ten ergeben, muss erwartet werden. In allen Ländern wurden und werden mindestens nach der Veröffentlichung der Ergebnisse von Large-Scale-Assessments (internationale, wie PISA oder TIMSS und nationalen, wie den Bildungstrends) Anstrengungen dazu unternommen, Gründe für solche Veränderungen zu identifizieren. Für die Auswahl geeigneter Maßnahmen sind diese Gründe bei positiven wie negativen Veränderungen von Interesse, wenngleich die Untersuchungen selbst keine kausalen Begründungen für Trends liefern können. Wie groß die Veränderung des wahren Wertes ist, sollte aber durch die Messungen des Bildungstrends angemessen abgeschätzt werden können. Der nächste Abschnitt wird diese Abschätzung an Hand der Leistungsverteilung auf Kompetenzstufen sowie der Mittelwerte auf der BiSta-Skala vornehmen, bevor danach Schlussfolgerungen für die Vergleichsarbeiten zusammengefasst werden.

5.3.1. Abschätzung der Stabilität für den Bildungstrends

Die Durchführungsbedingungen beim Bildungstrend sind deutlich strikter, als bei den Vergleichsarbeiten. Auf diese und weitere Unterschiede wurde schon im Abschnitt 1.2 eingegangen. Verwiesen sei hier insbesondere auf die Gegenüberstellung in der Tabelle 1.1. Die Konstruktion des Bildungstrends erfolgte mit dem Ziel, Messfehler durch stabil gehaltenen Durchführungsbedingungen so weit wie möglich zu reduzieren. Für die Vergleichsarbeiten sind alle diese Aspekte ungleich schwieriger umzusetzen. Vorteilhaft ist hier lediglich, dass es sich bei den Vergleichsarbeiten um eine Vollerhebung handelt. Nicht zuletzt aber wegen der professionellen Stichprobenziehung kann im Großen und Ganzen davon ausgegangen werden, dass die Erhebungen des Bildungstrends eine stabile Messung der Kompetenzstände erlauben. Das bedeutet: Wenn im Bildungstrend für eine Kompetenz eine Entwicklung von beispielsweise 18 BiSta-Punkten als statistisch signifikant ausgewiesen wird, kann mit großer Sicherheit davon ausgegangen werden, dass dies eine reale Entwicklung darstellt. Diese Entwicklung sollte sich auch in den Messungen der Vergleichsarbeiten zeigen. Natürlich muss hierfür das Konfidenzintervall beachtet werden und vor allem, dass eine im Bildungstrend im Abstand von 5 oder 6 Jahren gemessene Differenz nur in Anteilen davon zwischen zwei innerhalb nur eines Jahres aufeinander folgenden Messungen bei VERA erwartet werden kann. So wurden die Kompetenzen für Deutsch Lesen in der Sekundarstufe 2009 im ersten und 2015 im zweiten Zyklus erhoben, also in einem Abstand von 6 Jahren (Stanat et al., 2016). Für die beispielhafte Veränderung von 18 Punkten könnte man eine Veränderung von 3 Punkten pro Jahr angeben. Diese Veränderungen können natürlich auch nicht-linear verlaufen, zum

Beispiel wenn Interventionen zu einem bestimmten Zeitpunkt stattgefunden haben. Zudem findet die Messung der Kompetenzstände im Bildungstrend in der neunten Jahrgangsstufe statt und die der Vergleichsarbeiten in der achten. Eine Intervention Anfang der neunten Klasse würde sich nur im Bildungstrend abzeichnen können. Die durch den Bildungstrend angezeigten Veränderungen erlauben deshalb nur eine Abschätzung der Veränderungen des wahren Wertes, wie diese bei den Vergleichsarbeiten erwartet werden können.

Aus der Übersicht über die verschiedenen Messzeitpunkte des Bildungstrends und der Vergleichsarbeiten im Anhang (siehe Tabelle A.7) wird die Struktur der Zyklen beim Bildungstrend deutlich. Der erste Zyklus startete mit der Erhebung zu den sprachlichen Fächern Deutsch und erster Fremdsprache (Englisch und Französisch) in der Klasse 9 im Jahr 2009. Zwei Jahre später wurden die Kompetenzen für Deutsch und Mathematik in den vierten Klassen der Primarstufe erhoben. Den Abschluss bildete die Erhebung im Jahr 2012, welche die mathematischen und naturwissenschaftlichen Kompetenzen der Neuntklässler*innen fokussierte. 2015 startete der zweite Erhebungszyklus. Für 2021 beginnt der wegen der Corona-Pandemie um ein Jahr verzögerte dritte Zyklus. Die Vergleichsarbeiten finden hingegen jährlich statt. Für die Pilotierungen werden zusätzlich jährliche Testungen angesetzt. Mit der Pilotierung werden die VERA-Aufgaben für das Folgejahr pilotiert und normiert. Die Länder Deutschlands sind für diese Pilotierungen in zwei Gruppen zu je 8 Ländern aufgeteilt worden[4], um den organisatorischen Aufwand zu reduzieren. Jedes Land nimmt in jedem Jahr nur an der Pilotierung von VERA-8 oder VERA-3 teil[5]. Für Berlin und Brandenburg sind die Eintragungen in der Tabelle A.7 fett vorgenommen worden, wenn in dem betreffenden Jahr auch eine Pilotierung stattgefunden hat.

Die Ergebnisse von Kompetenzmessungen, ob im Bildungstrend oder bei den Vergleichsarbeiten, und speziell die Veränderungen zwischen zwei Messzeitpunkten lassen sich sowohl als veränderte Verteilung der Kompetenzstufen anzeigen, wie auch in einem Vergleich der BiSta-Mittelwerte der zwei Verteilungen. Solche Mittelwerte werden bei VERA zumindest für Berlin und Brandenburg nicht berichtet. Hier sollen beide Darstellungen gegenübergestellt werden. Die zentralen Befunde aus den Bildungstrends der ersten zwei Zyklen sind in Abbildung 5.3 für die Kompetenzstufenverteilungen des Landes Berlin dargestellt (Quellen für Deutsch und Mathematik der Primarstufe: Stanat et al., 2017b, S.8 Tab. 5.4web, S.14 Tab. 5.21web; für Deutsch und Englisch der Sekundarstufe: Stanat et al., 2016, S.209 Tab. 5.19,

[4] In die Gruppe der Länder, die mit Berlin und Brandenburg zusammen an den Pilotierungen teilnehmen gehören Bayern, Hessen, Mecklenburg-Vorpommern, Rheinland-Pfalz, Sachsen und Schleswig-Holstein.

[5] Für die Pilotierung der Domänen des Faches Französisch gelten spezifische Regelungen, da Französisch nicht in allen Ländern als erste Fremdsprache zugelassen ist.

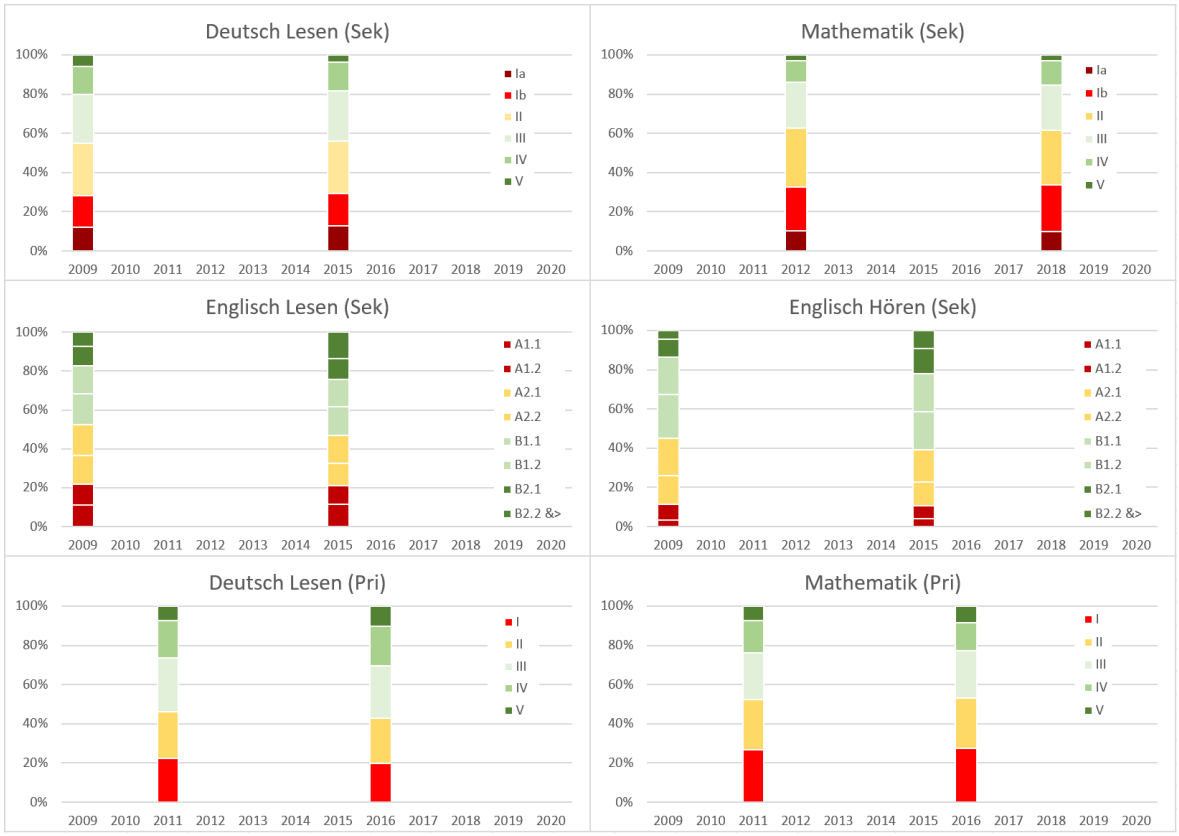


Abbildung 5.3.: Kompetenzstufenverteilungen der ersten zwei Zyklen des Bildungstrends für Berlin (Quellen im Text)

S.211 Tab. 5.20; für Mathematik der Sekundarstufe: Stanat et al., 2019b, S.22 Tab. 5.4web). In dieser und der folgenden Darstellung wurde die Zeitachse vollständig abgebildet, um dem Verhältnis jährlicher Messungen bei den Vergleichsarbeiten und dem nur alle 5 bzw. 6 Jahre stattfindendem Bildungstrend einen angemessenen Ausdruck zu verleihen. Die exakten Werte inkl. der berechneten Differenzen sind im Anhang A.6 gegenübergestellt (für die Primarstufe Tabellen A.10 und in der Sekundarstufe für Mathematik und Deutsch Tabelle A.8 und für Englisch Tabelle A.9).

Abbildung 5.3 zeigt für das Land Berlin, dass für die drei Domänen im Fach Deutsch in der Sekundarstufe die maximalen Veränderungen über alle Kompetenzstufen hinweg zwischen 2 und 3 Prozent liegen, mit Maxima bei den Stufen III oder V. Jeweils auf der obersten Kompetenzstufe *B2.2 und größer* liegen die maximalen Veränderungen im Fach Englisch für Lesen mit 6,2 und Hören mit 4,8 Prozent etwas höher. In Mathematik liegt der Wert bei nur 2,2 für die Sekundarstufe und bei 2,1 Prozent für den Primarbereich. Sonst findet sich für Deutsch im Primarbereich eine maximale Veränderung der Kompetenzstufenbelegung von 2,7 Prozent. Die Berichte zu den Bildungstrends weisen dabei lediglich für beide Englisch-Domänen

einzelne signifikante Differenzen aus, die in der Tabelle gekennzeichnet sind. Die Tabellen im Anhang sind äquivalent zu den Berichten im Bildungstrend durch Angaben für die Teilgruppe der Schülerinnen und Schüler ergänzt worden, die an Gymnasien unterrichtet werden. Hier sind die maximalen Änderungen der Belegungen einzelner Kompetenzstufen im Allgemeinen etwas größer, allerdings wird wegen der kleineren Stichprobe auch der Standardmessfehler und damit das Konfidenzintervall größer, so dass auch hier lediglich im Fach Englisch wenige signifikante Verbesserungen ausgewiesen wurden. Bis auf die Verbesserungen im Fach Englisch, die offensichtlich darauf zurückzuführen sind, dass mehr Schüler*innen statt einer mittleren eine hohe Kompetenzstufe erreicht haben, lassen sich für das Land Berlin keine bedeutsamen Veränderungen zwischen den Erhebungen der zwei Zyklen feststellen.

Für die Mittelwerte der Kompetenzverteilungen über der Metrik der Bildungsstandards werden in den Berichten zum Bildungstrends ebenso Aussagen gemacht, auf die hier durch eine zusammenfassende graphische Darstellung im Text und eine Tabelle im Anhang rekurriert wird (Quellen für Deutsch und Mathematik der Primarstufe: Stanat et al., 2017a, S.159 Abb. 6.4, S.167 Abb. 6.16; für Deutsch der Sekundarstufe: Stanat et al., 2016, S.347-351 Abb. 6.7, 6.9, 6.11, S.353-535 Abb. 6.13, 6.14, 6.15; für Englisch der Sekundarstufe: Stanat et al., 2016, S.368, 370 Abb. 6.20, 6.22, S.373 Abb. 6.24, 6.25; für Mathematik der Sekundarstufe: Stanat et al., 2019a, S.207 Abb. 6.4, S210 Abb. 6.6).

Die Mittelwerte sind in der Abbildung 5.4 für alle Domänen abgetragen, die auch im VERA-Kontext durchgehend oder in mehreren Jahren überprüft wurden. Zudem sind die in der Tabelle explizierten Ergebnisse für die Schüler*innen an Gymnasien auch hier ergänzt. Erwartungsgemäß bestätigen sich die aus der Analyse der Kompetenzstufen ergebenen Erkenntnisse: Es gibt marginale Veränderungen für Deutsch in der Sekundarstufe, wie auch für beide Fächer der Primarstufe, keine Veränderungen für Mathematik in der Sekundarstufe und sichtbare Verbesserungen für beide Domänen des Faches Englisch. Als signifikant erweisen sich die Verbesserungen in Englisch lediglich für die Veränderungen des Gesamtwertes, nicht für die Gymnasien allein. Die Veränderungen sind in der Sekundarstufe vom Betrag her in Mathematik tatsächlich Null, schwanken für die Deutsch-Domänen zwischen -6 im Lesen und +5 bei Orthographie und in Englisch zwischen +15 Punkte beim Lesen und +20 für Hören, jeweils als Veränderung über einen Zeitraum von 6 Jahren. In der Grundschule findet sich mit +8 Punkten beim Lesen die größte positive und mit -5 Punkten die größte negative Veränderung über einen Zeitbereich von 5 Jahren. Für die auch bei VERA durchgängig geprüften Domänen werden in der Tabelle die Ergebnisse für den Bildungstrend im oberen Teil

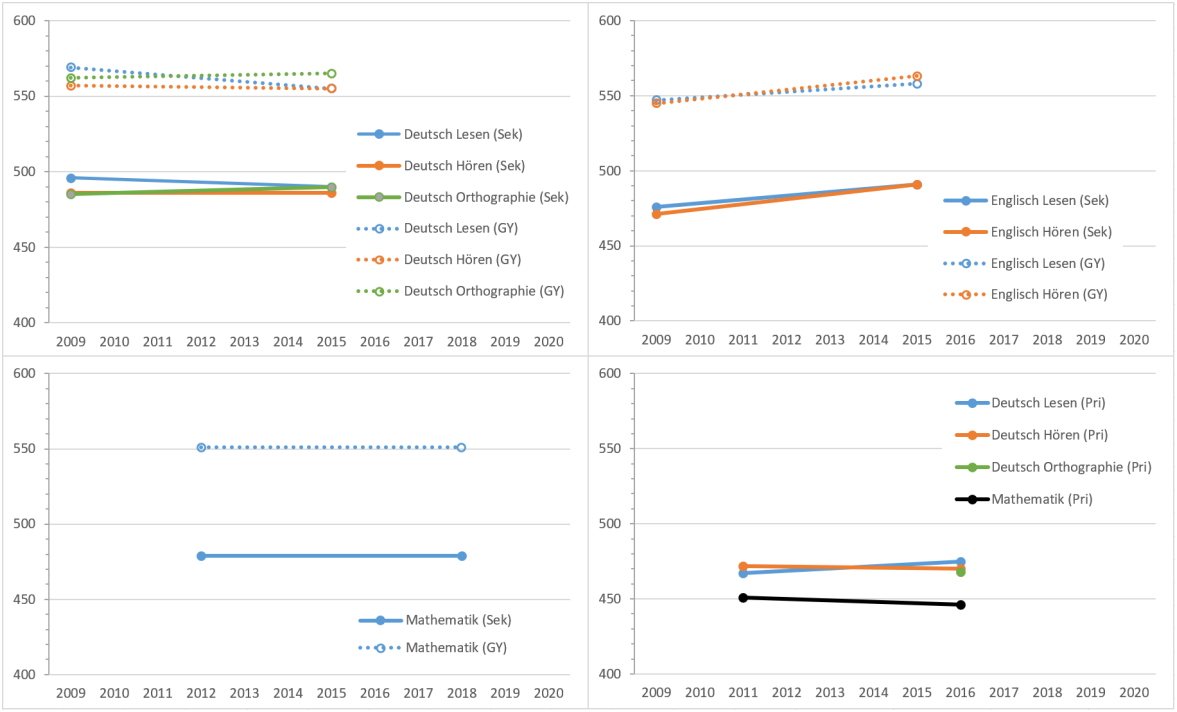


Abbildung 5.4.: Mittelwerte auf der Skala der Bildungsstandards der ersten zwei Zyklen des Bildungstrends für Berlin

der Tabelle 5.1 numerisch dargestellt. Die Veränderungen werden hier in der Zeile *abs. pro Jahr* als relative Differenz der absoluten Veränderung innerhalb eines Jahres wiedergegeben[6].

Die letzten Darstellung haben einen Eindruck darüber vermittelt, welche Veränderungen sich beim Bildungstrend über einen Zeitraum von 5 bzw. 6 Jahren üblich ergeben. Damit werden die Größenordnungen erwartbarer Veränderungen bei den jährlichen Vergleichsarbeiten deutlich. Der folgende Abschnitt soll dies gegenüberstellen.

5.3.2. Schlussfolgerungen für die Vergleichsarbeiten

Auch unter der Annahme, dass Entwicklungen nicht stetig stattfinden, sollten die jährlich gemessenen Variationen der BiSta-Mittelwerte wie auch die Veränderungen der Besetzung von Kompetenzstufen bei den Vergleichsarbeiten kleiner sein, als äquivalente Veränderungen bei der Messung im Rahmen der Bildungsstandards in einem Turnus von 5 bzw. 6 Jahren, wie im vorherigen Abschnitt dargestellt. Allerdings lassen sich auch einige Gründe aufführen, warum die bei den Vergleichsarbeiten ergebende Differenzen von diesen Erwartungen abweichen können. Einerseits solche, die sich in einer veränderten Streuung niederschlagen

[6]Hierbei wird der absolute Betrag der Veränderung durch die Zahl der Jahre zwischen den Messungen geteilt. Für die Berechnung der Veränderungen bei den Vergleichsarbeiten im unteren Teil der Tabelle, wird der absolute Betrag der jährlichen Veränderung gemittelt.

Tabelle 5.1.: Gegenüberstellung der Veränderungen der erreichten Kompetenzen beim Bildungstrend und bei den Vergleichsarbeiten

	Deutsch Lesen[a]			Mathematik		Englisch Lesen		Englisch Hören[a]	
VERA	3	8	8 Gy	8	8 Gy	8	8 Gy	8	8 Gy
Messungen beim Bildungstrend (Originalwerte und Differenz)									
Zyklus 1	467	496	569	479	551	476	547	471	545
Zyklus 2	475	490	550	479	551	491	558	491	563
Jahre	5	6	6	6	6	6	6	6	6
Differenz[b]	8	-6	-14	0	0	15	11	20	18
abs. pro Jahr[c]	1,6	1,0	2,3	0,0	0,0	2,5	1,8	3,3	3,0
Messungen bei VERA (Differenzen zum Vorjahr)									
2011	31	31	34	4	4	-44	-34		
2012	-3	-7	-27	14	8	20	31	-4	-14
2013	-28	1	23	-31	-36	-29	-37	26	38
2014	27	-13	-37	32	32	-21	-28	-38	-39
2015	17	21	30	16	7	52	53	21	8
2016	8	-25	-31	-35	-21	8	31	22	37
2017	-35	23	38	17	9	-36	-58	12	9
2018	5	-42	-26	-17	-14	54	62	-8	-14
2019	-3	14	9	3	-4	-8	5	14	33
2020		0	9	32	35	-9	-12	-1	-3
Mw Differenz[d]	2,1	0,3	2,2	3,5	2,0	-1,3	1,3	4,9	6,1
Sd Differenz[d]	21,1	21,8	28,1	22,7	20,9	32,8	39,2	19,0	25,0
abs. pro Jahr[c]	17,4	17,7	26,4	20,1	17,0	28,1	35,1	16,2	21,7

[a] 2011 wurde die Domäne Englisch Hörverstehen nicht getestet. 2020 fiel die gesamte VERA-3-Testung pandemiebedingt aus.
[b] Differenz zwischen den zwei vorliegenden Messungen.
[c] relative Differenz der absoluten Veränderungen, Punkte pro Jahr.
[d] Mittelwert und Standardabweichung der Differenzen über die Jahre.

könnten:

- Die Besetzung einzelner Kompetenzstufen unterliegt wegen der notwendigen Verwendung von WLE-Schätzern gegenüber plausible values (PV) beim Bildungstrend stärkeren Schwankungen, wie im vorhergehende Kapitel dieser Arbeit schon erwähnt.

- Standardfehler und somit Konfidenzintervalle sollten wegen der Vollerhebung bei VERA im Gegensatz zum Stichprobendesign beim Bildungstrend deutlich kleiner sein.

- Während beim Bildungstrend mit dem Rotationsdesign eine weite Streuung der Inhalte innerhalb einer Domäne erfolgt, ist dies bei VERA, wegen der Verwendung von nur ein oder zwei verschiedenen, zudem fest zugeteilten Testheften anders. Dass der Test die Inhalte repräsentativ operationalisiert ist für VERA demnach deutlich schlechter realisiert (vergleiche dazu auch die Abbildung 3.4).

Weitere Einschränkungen bei der Übertragung von Veränderungen beim Bildungstrend auf die Vergleichsarbeiten führen vermutlich zu Verzerrungen, also solchen Veränderungen, die sich im Mittel nicht zu Null summieren. Das sind zum Beispiel:

- Durch den Ausschluss von privaten Schulen bei den Vergleichsarbeiten[7] ergeben sich ggf. Verschiebungen bei den Ergebnissen zwischen Bildungstrend und Vergleichsarbeiten, vermutlich zugunsten des Bildungstrends. Dies könnte auch einen Einfluss auf die Messung der Veränderung haben, wenn sich die Anteile von Schulen in privater Trägerschaft deutlich ändern.

- Zudem müssen Unterschiede bei der Messung der Leistung wegen der bei den Vergleichsarbeiten deutlich weniger restriktiven Durchführungsbedingungen vermutet werden. Nicht zuletzt sind hier konkrete Beeinflussungen durch Lehrkräfte möglich bzw. zu erwarten, vermutlich zu Gunsten der Ergebnisse der Vergleichsarbeiten. Auch diese Einflüsse sind nicht notwendig über die Jahre stabil.

- Die für VERA spezifische Form der Verlinkung birgt die Möglichkeit gewisser Instabilitäten. Während durch die Wiederverwendung eines großen Teils der Aufgaben beim Bildungstrend die Verlinkung einfach gegeben ist, spielt sie für die Verortung der VERA-Instrumente auf der BiSta-Metrik eine essentielle Rolle. In Verbindung mit der Tatsache,

[7] Einrichtungen privater Träger dürfen durchaus an den Vergleichsarbeiten teilnehmen und werden dabei auch identisch unterstützt (Anmeldung, Informationsveranstaltungen, Druck der Unterlagen und Begleitmaterialien, Eingabeportal, Rückmeldungen, begleitende Hotline). Da deren Teilnahme aber nur freiwillig erfolgt, werden deren Ergebnisse in keine Berechnung mit einbezogen. Tatsächlich erhalten diese privaten Schulen deshalb in den Rückmeldungen die Vergleichswerte aller öffentlichen Schulen.

dass a) die Pilotierungen von den Ländern selbst und offensichtlich in unterschiedlicher Form organisiert werden, b) nur jeweils 8 Länder und nicht in jedem Fall alle 16 beteiligt sind sowie c) nach der Pilotierung kein zweiter Einsatz der späteren Testhefte zur Normierung erfolgt, sondern Testblöcke aus der Pilotierung als in sich stabil angesehen werden, kann vermutet werden, dass hier Quellen von Instabilitäten existieren.

Das könnten Gründe dafür sein, dass sich die Messung der Kompetenzen insbesondere im Rahmen der Vergleichsarbeiten als instabil erweist. Und zwar sowohl zwischen verschiedenen VERA-Messungen wie auch zwischen VERA und dem Bildungstrend. Trotz dieser Einschränkungen kann aber erwartet werden, dass die im Bildungstrend gefundenen Veränderungen einen groben Rahmen für solche bei VERA abstecken. Sofern sich Rahmenbedingungen der Durchführung von VERA oder deren Einbettung im Land nicht deutlich verändert haben, sind zudem einige der Einflüsse für den Vergleich von zwei VERA-Messungen als Trendbetrachtung nicht relevant.

Der untere Teil der Tabelle 5.1 stellt die Veränderungen aus den Vergleichsarbeiten denen des Bildungstrends gegenüber, wobei in den Zeilen jeweils die Veränderungen zum Vorjahr abgetragen sind. Für jede Domäne findet sich ein ähnliches Bild: Alle Messwerte streuen mit einer Standardabweichung von 20 bis 30 Punkten um einen Mittelwert, der über die Jahre bei ca. Null liegt. Es ist also keine Tendenz auszumachen. Die Messungen streuen aber erheblich um den Mittelwert. Zieht man den absoluten Wert für die Mittlung heran, so ergibt sich bei VERA eine jährlich Veränderung, die für die verschiedenen Domänen zwischen 16,2 und 35,1 Punkten pro Jahr liegt und damit vielfach höher als beim Bildungstrend[8].

Als Resultat muss die zuvor formulierte Erwartung, nach der die Veränderung einer mittleren VERA-Leistung für das Land Berlin zwischen zwei Testzeitpunkten in einem angemessenen Verhältnis zu den Veränderungen in der gleichen Domäne beim das VERA-Zeitintervall einschließenden Bildungstrend stehen sollte, verworfen werden. Die überraschend großen Veränderungen zwischen den jährlichen VERA-Messungen sollen folgend näher betrachtet werden, um die Validität der Trendbetrachtung und damit jeder Einzelmessung zu untersuchen. Im ersten von zwei Abschnitten werden die Ergebnisse von Testungen im Rahmen der standardisierten VERA-Durchgänge aus verschiedenen Jahren gegenübergestellt. Für den zweiten Abschnitt werden zwei ISQ-Studien (Graf et al., 2016; Vettorazzi & Harych, 2019) der letzten Jahre, die sich auf Daten der Vergleichsarbeiten stützen auf Aspekte der Stabilität hin untersucht.

[8] Für Mathematik lässt sich dieser Faktor nicht berechnen, weil die absolute Veränderung von im Mittel 20 Punkten pro Jahr bei VERA einer von Null beim Bildungstrend gegenübersteht.

5.4. Untersuchung der Stabilität im Rahmen standardisierter VERA-Tests

Die folgenden Betrachtungen einzelner Artefakte stellen keine zufällige Auswahl dar, sondern fokussieren wiederholt auftretende Besonderheiten an ausgezeichneten Stellen, vermutlich denen größerer Abweichungen vom Erwarteten. Zur Bewertung der Bedeutung dieser Artefakte ist es wichtig zu klären: Sind die Artefakte nur Zeichen von erwartbaren statistischen Schwankungen, also von „Rauschen“ in den Daten oder doch Verzerrung? Die Darstellungen nehmen im Wesentlichen Bezug auf veröffentlichte Ergebnisse der Vergleichsarbeiten für das Land Berlin, die den Berichten bis zu den Jahrgängen 2015/16 auf der Webseite des ISQ (www.isq-bb.de) zu entnehmen sind. Im Einzelnen werden diese Daten durch solche ergänzt, die durch Anfragen im Berliner Landesparlament öffentlich wurden.

Dargestellt wird zuerst eine unerwartet große Verschlechterung von Mathematikergebnissen bei VERA-8 im Jahr 2016 zur Vorjahreserhebung. Diese findet sich auch in der von Köller et al. (2020) berichteten Abbildung 5.2[9]. Der zweite Fall ist eine erhebliche Vergrößerung des Anteils von Schüler*innen, die 2016 beim Leseverstehen in der Primarstufe die Kompetenzstufe V erreicht haben. Im dritten Abschnitt werden die Ergebnisse, die Sekundarschüler*innen der achten Jahrgangsstufe 2014 im Leseverstehen bei den Vergleichsarbeiten zeigten, denen der gleichen Kohorte ein Jahr später beim Bildungstrend gegenübergestellt.

5.4.1. Verschlechterung der Mathematikergebnisse in Berlin von 2015 zu 2016 bei VERA-8

Die Abbildung 5.4 zeigt in der Graphik unten links für alle Berliner Schüler*innen der Sekundarstufe I und zusätzlich noch nur für alle Gymnasiast*innen unter diesen, die Mathematikleistungen zu zwei Messzeitpunkten im Rahmen der Erhebungen zum Bildungstrend (Stanat et al., 2019a, S.207 Abb. 6.4, S210 Abb. 6.6). Die Ergebnisse dieser höchsten wissenschaftlichen Standards genügenden Untersuchung weisen für die Jahre 2012 und 2018 für das Land Berlin exakt den gleichen Wert aus. Es hat also im Land Berlin bis auf zufällige Schwankungen keine Veränderungen bei den Mathematik-Leistungen von Schüler*innen der neunten Jahrgangsstufe gegeben. Auch, wenn man sich nur die Ergebnisse der Gymnasien betrachtet, zeigt sich ein identisches Ergebnis. Ein ebenso äquivalentes Bild findet sich für

[9] Die Darstellung bei Köller et al. (2020) zeigt die Verschlechterung allerdings in anderer Form und zwar als den Anteil an Schüler*innen, die den Mindeststandard nicht und damit nur die Kompetenzstufe I erreichen. Dieser Anteil steigt von 2015 nach 2016 entsprechend an.

die Verteilung der Kompetenzstufen (siehe Abbildung 5.3, oben rechts).

Die Vergleichsarbeiten messen das äquivalente Konstrukt, denn das Instrument wird vom IQB mit dem in Abschnitt 3.6 dargestellten Verfahren entwickelt und misst damit auf der identischen Metrik. Für die Reliabilität der Messung ist vorteilhaft, dass die Leistungen der Achtklässler*innen im Rahmen der Vergleichsarbeiten als Vollerhebung erfasst werden und Stichprobenfehler die Güte der Messung damit nicht beeinträchtigen können. Allerdings wird die Erhebung nicht durch Testleiter*innen administriert und auch die Kodierung der Schülerantworten erfolgt zwar auf der Basis eines detaillierten Manuals, allerdings ebenso durch diesbezüglich ungeschulte Lehrkräfte. Dem früheren Messzeitpunkt in der Jahrgangsstufe 8 kann zugeschrieben werden, dass der Anteil an Schüler*innen, deren Leistungen auf der untersten Kompetenzstufe[10] liegen bei den Vergleichsarbeiten 2012 bei 58% liegen und bei der fast zeitgleichen Messung des Bildungstrends bei den Schüler*innen eines Jahrgangs darüber bei nur noch 32,7%. In der zweiten Welle der Erhebungen zum Bildungstrend 2018 stehen sehr ähnlich 68% bei den Vergleichsarbeiten 33,9% gegenüber (ebenda).

Die Graphik 5.5 führt die Ergebnisse der unterschiedlichen Messungen zusammen und macht sie damit einfach vergleichbar (obere Graphik). Um die Ergebnisse direkt mit denen des Bildungstrends vergleichen zu können, wurden äquivalent zu diesem die Ergebnisse aller Schüler*innen um die Ergebnisse nur der Schüler*innen an Gymnasien ergänzt. Alle Ergebnisse werden relativ zum Ergebnis des Jahres 2012 dargestellt, dem Zeitpunkt der Erhebung des ersten Bildungstrends.

Für Berlin (rote Linien) fällt zuerst der Gleichlauf für alle Schüler*innen (untere Linie) und für Gymnasiast*innen (obere Linie) auf. Der Anteil der Gymnasiast*innen ist dabei über die Jahre relativ stabil zwischen 45,8% (2013) und 48,9% (2012) (siehe untere Graphik). Zum Vergleich wurden die Ergebnisse der zwei Messungen zum Bildungstrend in der gleichen Graphik abgebildet. Erwartungsgemäß ist das Konfidenzintervall beim Bildungstrend auf Grund der Stichprobe deutlich größer. Insbesondere für die Mathematik, bei der es weder in der Gesamtstichprobe noch bei den Gymnasiast*innen eine Veränderung zwischen den zwei Messzeitpunkten des Bildungstrends gegeben hat, fallen die Schwankungen bei den Vergleichsarbeiten ins Auge. Für die Berliner Werte der Vergleichsarbeiten im hier berichteten Zeitraum von 2010 bis 2020 findet sich eine Spannweite von etwa einem Schuljahr (48 Punkte). Im Jahr 2013 wird ein Tiefpunkt erreicht, der zwei Jahre später erreichte Hochpunkt wird im Jahr 2020 eingestellt. Die größte Veränderung zwischen zwei aufeinanderfolgenden

[10] das meint die Kompetenzstufe I und bezieht ggf. geteilte Stufen Ia und Ib mit ein.

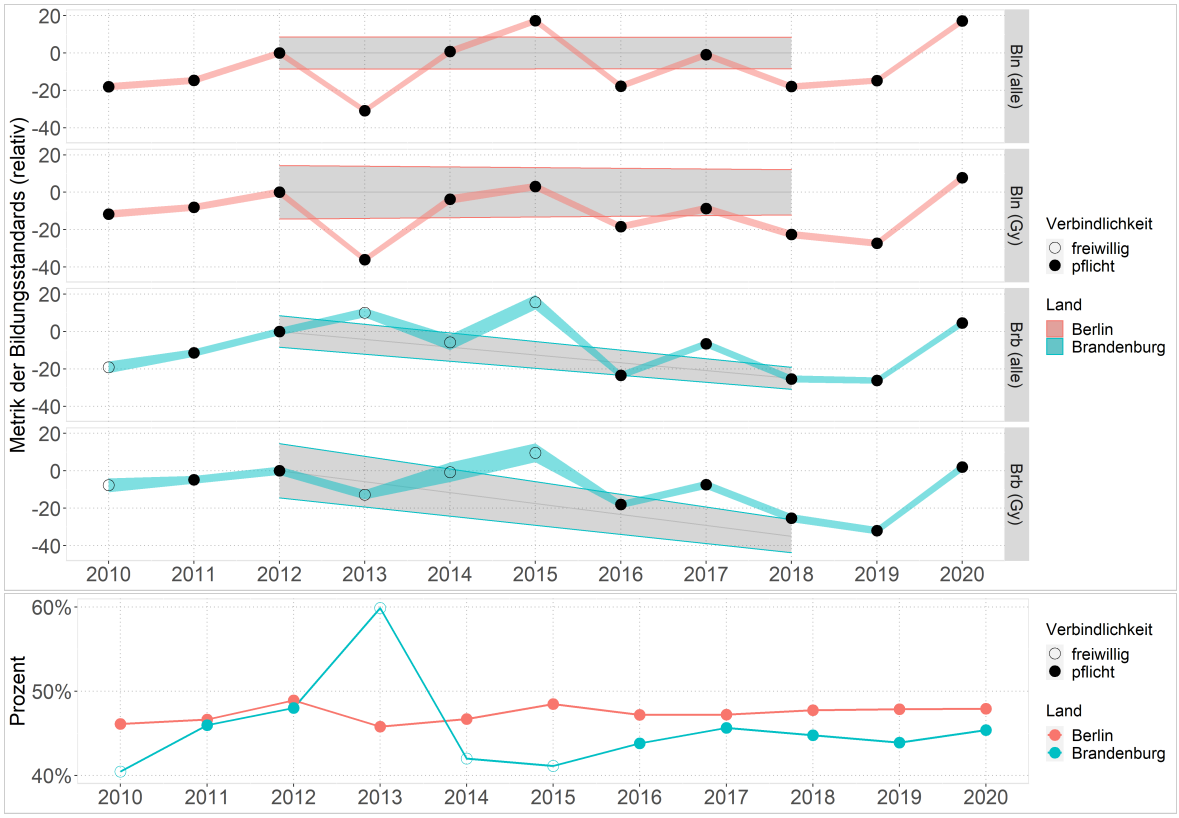


Abbildung 5.5.: Oben: Gegenüberstellung der Ergebnisse von Bildungstrend und VERA-8 für Berlin und Brandenburg, jeweils für alle Schulen und für Gymnasien, für 2010 bis 2020, wobei die Ergebnisse relativ zum Ergebnis von 2012 dargestellt sind, dem Zeitpunkt der ersten Erhebung des Bildungstrends. Unten: Anteil der Gymnasiast*innen bei den Vergleichsarbeiten.

Jahren findet sich von 2015 auf 2016 mit einem Verlust von 35 Punkten.

Um diese Veränderungen in der Größenordnung einschätzen zu können, findet man bei Stanat et al. (2019a, S. 201) für den erwarteten Kompetenzzugewinn im Fach Mathematik innerhalb eines Schuljahres am Ende der Sekundarstufe I einen Wert von etwa 50 Punkten. Hierbei beziehen sich die Autor*innen auf Schätzungen aus den Normierungsstudien des Bildungstrends. Da sie hierbei den Zugewinn von der Jahrgangsstufe 9 nach 10 untersucht haben, werden jene Schüler*innen ausgeschlossen, die nach ihrem Hauptschulabschluss am Ende der neunten Klasse die Schule verlassen. Beim Bildungstrend wird zudem die Spannweite zwischen den Ländermittelwerten in Deutschland mit 70 Punkten als in etwa eineinhalb Schuljahren angegeben (Stanat et al., 2019a, S. 207). Sind also Veränderungen zwischen zwei Jahren in Größenordnungen von bis zu 35 Punkten, wie sie sich bei den Vergleichsarbeiten darstellen plausibel? Wären Messungen des Bildungstrends dann nicht ein Stück weit zufällig? Das Land Berlin könnte mit einer Veränderung von 35 Punkten in einem Jahr im Mittelfeld der 16 Länder stehen und ein Jahr später Schlusslicht sein. Folgende unterschiedlichen Begründungen

könnten aus Sicht des Autors solche Veränderungen erklären:

- Die Veränderungen sind normale, zwischen den Kohorten zu erwartende Leistungsschwankungen, also zufällig.
- Die Veränderungen lassen sich auf Umgebungsvariablen zurückführen, welche die Leistungen entweder direkt beeinflussen, wie die Einführung eines neuen Schulbuchs oder eine Weiterbildungsinitiative oder auch indirekt, wie zum Beispiel die Veränderung der Kohorte durch besonders starke temporäre Zuwanderung.
- Die Instrumente sind unzureichend miteinander verlinkt, so dass die sich in den Daten widerspiegelnden Schwankungen nicht auf reale Schwankungen zurückzuführen sind.

Gegen die erste Annahme spricht der auch wegen der Vollerhebung sehr kleine Standardfehler, der sich im kleinen Konfidenzintervall deutlich zeigt. Zur grundlegenden Prüfung der zwei anderen Annahmen wären spezifische Untersuchungsdesigns notwendig. Im Folgenden können aber vorliegende Daten genutzt werden, um die Annahmen zu konkretisieren.

Als beeinflussende Umgebungsvariablen ließen sich natürlich zu viele finden, die alle schwerlich zu überprüfen sind. Zudem wären auch alle Interaktionen mehrerer solcher Variablen zu untersuchen. Leicht können allerdings erhobene Einflussgrößen untersucht werden. So ist der sehr äquivalente Verlauf der Schätzungen einerseits für die Gesamtstichprobe und andererseits für die der Gymnasiast*innen allein schon interessant. Berlin und Brandenburg sind zwei strukturell sehr unterschiedliche Länder. In den regionalen Bildungsberichten (vergleiche zuletzt Wendt et al., 2017) werden diese Unterschiede deutlich. Der Bildungstrend weist deshalb auch erwartbar unterschiedliche Ergebnis für beide Länder aus. Dabei sind die Ergebnisse für das Land Brandenburg durchgehend besser, als für das Land Berlin. Bei den in der Darstellung 5.5 auch für Brandenburg abgebildeten Ergebnissen über die Jahre fällt auch hier der sehr ähnliche Verlauf sofort ins Auge. Es ist allerdings zu beachten, dass die Teilnahme am Mathematiktest bei VERA-8 in Brandenburg nicht durchgehend verpflichtend war. Konkret war die Teilnahme im betrachteten Zeitraum in vier Jahren freiwillig: 2010 und 2013 bis 2015 (siehe auch leere vs. ausgefüllte Markierungen in Abbildung 5.5). Nimmt man diese vier Zeitpunkte bei einer ersten Betrachtung aus, dann liegt Brandenburg 2011 noch 34 Punkte vor Berlin und ein Jahr später noch 31 Punkte. Der Abstand sinkt dann von 2016 mit 25 Punkten bis 2020 kontinuierlich bis auf 18 Punkte ab. Drei der Messzeitpunkte mit freiwilliger Brandenburger Beteiligung lassen sich in diesen Verlauf der Abstände gut einordnen. Nur 2013 fällt hier heraus; hier schneidet Brandenburg deutlich besser ab, als man es dem

Verlauf nach erwarten würde. Wie man im unteren Teil der Graphik 5.5 allerdings sehen kann, fällt dieses Jahr durch einen unerwartet hohen Anteil von teilnehmenden Gymnaiast*innen im Land Brandenburg auf. Während dieser Anteil sonst bei rund 45% und selbst bei den freiwilligen Teilnahmen 2010, 2014 und 2015 nur leicht darunter liegt, kommen 2013 fast 60% der an den Vergleichsarbeiten für Mathematik beteiligten Brandenburger*innen aus Gymnasien. Während also über die Jahre die Messwerte aus den VERA-Erhebungen nicht unerheblich schwanken, entwickeln sich die Differenzen zwischen den Ergebnissen der zwei Länder Berlin und Brandenburg auffallend stabil. Zudem lässt sich der abnehmende Leistungsabstand zwischen Berlin und Brandenburg ebenso auf der Ebene des Bildungstrends finden. Der Unterschied von 39 Punkten aus dem Jahr 2012 wird bis 2018 auf nur 14 Punkte mehr als halbiert, und er bewegt sich damit auf einem ähnlichen Niveau wie bei VERA. Die insgesamt beobachteten Schwankungen, verändern sich zwischen den Schulformen und zwischen den Ländern nicht. Wenn eine Umgebungsvariable die Schwankungen der VERA-Messungen erklären soll, so müsste diese in den Schulformen und in beiden Ländern äquivalent wirken, was sehr unwahrscheinlich ist.

So bleibt die Vermutung zu prüfen, ob Schwankungen auf eine unzureichende Verlinkung der verwendeten Instrumente zurückzuführen sind. Für die Gymnasien und die nicht-gymnasialen Schulformen werden unterschiedliche, wenn auch nicht vollständig disjunkte Testhefte eingesetzt (siehe auch Abschnitt 3.5 für die Konstruktion der Testhefte und A.3 für die Auswahl der jährlich eingesetzten Testheftversionen für das Land Berlin). Dass rund ein Drittel der Aufgaben beider Testheftversionen identisch sind, sorgt trotz testheftweiser Skalierung für eine Verlinkung über die festen Itemparameter dieses Drittels der Aufgaben. Dies erklärt vermutlich die stabile Differenz zwischen den Werten der Schulformen zu jedem Messzeitpunkt. Einziger Ausreißer bei den parallelen Verläufen über die Jahre ist das Jahr 2013. Der Tabelle A.3 kann man entnehmen, dass nur in diesem Jahr nicht die einfachste Version A den nicht-gymnasialen Schulen und die mittelschwere Version B den Gymnasien zugeordnet wurden, sondern jede Schulform mit dem schwereren Testheft ausgestattet wurde. So erhielten in diesem (und nur in diesem) Jahr die nicht-gymnasialen Schulen die Version B und die Gymnasien die schwere Version C. Die freiwillige Teilnahme Brandenburger Schüler*innen, verbunden mit dem in diesem Jahr überproportionalem Anteil von Schüler*innen aus Gymnasien erklären, dass die Parallelität über die Jahre allein hier aussetzt. Die Vermutung, dass die direkte Verlinkung zwischen Instrumenten, wie sie bei den Testheftversionen eines Jahres vorliegt, zu einer stabileren Verlinkung respektive Metrik führt, als die nur mittelbare Verlin-

kung zwischen den Jahren, wird durch diese Daten gestützt. *Mittelbare Verlinkung* beschreibt, die in Abbildung 3.4 erkennbare Linking-Brücke, welche die VERA-Aufgaben mit den ausgewählten Anker-Aufgaben verlinken. Zwei VERA-Aufgaben-Sets sind dann über zwei solcher Brücken verknüpft, was hier als mittelbare Verlinkung bezeichnet wird.

Zuletzt soll noch das Verhältnis der Ergebnisse zwischen dem Bildungstrend und denen der Vergleichsarbeiten untersucht werden. Die gemeinsame Metrik gestattet eine solche Gegenüberstellung. Wenn das IQB aus der Normierung einen Zuwachs von der Klassenstufe 9 zur 10 von etwa 50 Punkten ansetzt, kann man ähnliches für den Zuwachs von der Jahrgangsstufe 8, gemessen Ende Februar mit VERA-8, zur Jahrgangsstufe 9, gemessen beim Bildungstrend im Mai, ansetzen. Wegen der großen, bisher nicht erklärbaren Schwankungen bei den Vergleichsarbeiten, soll hier zuerst für das Land Berlin der beim Bildungstrend über die 6 Jahre stabile Wert von 479 dem Mittelwert aller VERA-Messungen gegenübergestellt werden. Das sind alle Messungen an und zwischen den zwei Messzeitpunkten des Bildungstrends und jeweils zwei davor und danach. Der sich ergebende Wert liegt zwar erwartungsgemäß unter dem des Bildungstrends, aber nicht im erwarteten Maße. Für Brandenburg wird eine solche Gegenüberstellung komplizierter, weil der Trend deutlich nach unten zeigt, hier also über die Jahre keine stabilen Mathematikleistungen angenommen werden können. Wenn man aber die VERA-Ergebnisse jeweils um den Zeitpunkt der zwei Erhebungen des Bildungstrends 2012 und 2018 mittelt, finden sich auch hier nur kleine Unterschiede zwischen VERA und Bildungstrend[11]. Auch in Brandenburg überschätzen die Ergebnisse der Vergleichsarbeiten im Vergleich zum Bildungstrend die Mathematikergebnisse deutlich.

Dass die Verschlechterungen der bei VERA gemessenen Mathematikleistungen auf echte Leistungsunterschiede zurückzuführen ist, scheint damit eher unwahrscheinlich. So sehr wie die Ergebnisse bei den Vergleichsarbeiten auch schwanken, so stabil ist dabei der Abstand zwischen den zwei so unterschiedlichen Länder. Man erkennt sogar die Abnehmende Differenz zwischen Berlin und Brandenburg beim Bildungstrend in der Annäherung der zwei Kurven für Berlin und Brandenburg bei VERA.

[11] Für diese Abschätzung wurden die VERA-Mittelwerte wieder aus dem Jahr der Bildungstrenderhebung sowie jeweils der zwei Jahre davor und danach gemittelt, also einmal jene von 2010 bis 2014 und dann von 2016 bis 2020. Dies ist eine sehr grobe Schätzung. Auch, weil sie über die schon hier offensichtlich schwer erklärbaren Schwankungen bei VERA mittelt. Die Tendenz ist aber deutlich

5.4.2. Steigerung des Anteils an Schüler*innen in der Kompetenzstufe V bei Deutsch Lesen in VERA-3

Seit das Land Berlin sich an den Vergleichsarbeiten in der Grundschule beteiligt hat, ist der Test des Leseverständnisses obligatorischer Bestandteil. Mit der Übernahme der Testentwicklung durch das IQB im Jahr 2010, erhalten die Lehrkräfte zudem eine klassenbezogene Rückmeldung über die Verteilung der Kompetenzstufen für die Domäne Deutsch Lesen. Diese Verteilung kann für das Land Berlin den Ergebnisberichten für das Jahr 2015 (Vettorazzi et al., 2015) und 2016 (Vettorazzi et al., 2017a) entnommen werden. Die Graphik 5.6 stellt oben links diese Ergebnisse und die vorhergehender Jahrgänge direkt gegenüber. Diskutiert wurde insbesondere nach der Durchführung 2016 der gegenüber dem Vorjahr 10%-ige Aufwuchs des Anteils von Schüler*innen, welche die Kompetenzstufe V erreichten, denn dieser findet sich in ähnlicher Form auch in anderen Ländern[12]. Die Ergebnisse für das Land Brandenburg (Vettorazzi et al., 2016; Vettorazzi et al., 2017b) zeigen ein ähnliches Bild. Der Anteil der Schülerinnen und Schüler mit Leistungen der Kompetenzstufe V steigerte sich um 12,4% (siehe Graphik 5.6, oben rechts). Eine solche Steigerung des Anteils von Schüler*innen auf der höchsten Kompetenzstufe wurde von Vertreter*innen verschiedener auswertender Einrichtungen als unerwartet stark bewertet. Das die Steigerung in mehreren Ländern auffällig war, wenn auch in unterschiedlicher Ausprägung, lässt einen überregionalen Einfluss vermuten. Spezifisch für Berlin und Brandenburg war hingegen die gleichzeitige Steigerung des Anteils der Schülerinnen und Schüler, welche nur die Kompetenzstufe I erreichten um 6,6% in Berlin und 5,2% in Brandenburg.

Für eine nähere Untersuchung der Ergebnisunterschiede werden zuerst die Testinstrumente in der Darstellung 5.7 gegenübergestellt. Unten sind für das Testheft des Jahres 2015 die 21 Itemparameter als gelbe Dreiecke[13] auf der Metrik der Bildungsstandards angeordnet, darunter die 22 Positionen möglicher Personenparameter. Äquivalent sind oben die entsprechenden Parameter des Testhefts aus dem Jahr 2016 mit 19 Items und 20 möglichen Positionen der Personenparameter abgebildet. Auf der mittleren Achse sind zusätzlich die Kompetenzstufen farblich markiert, so dass sich die Verhältnisse leicht interpretieren lassen. Wie eng die Lage der Personenparameter mit jener der Itemparameter verknüpft ist, wird in Abschnitt 4.1.2

[12] Hier kann leider auf keine Untersuchung aus anderen Ländern verwiesen werden. Zudem wurden offensichtlich Berechnungen einzelner Länder mit dem Ziel einer kontinuierlichen Berichterstattung plausibilisiert. Dass der Effekt länderübergreifend beobachtet wurde erwähnt allerdings auch Weirich (2016).

[13] Tatsächlich gibt es 22 Items. Das schwierigste wurde in den Testheften für die Länder Berlin und Brandenburg aus inhaltlichen Erwägungen nicht eingesetzt und ist deshalb hier nur ergänzend als leeres Dreieck dargestellt. Alle folgenden Berechnungen beziehen sich auf genau dieses Testheft mit nur 21 Items.

Abbildung 5.6.: Verteilung der Kompetenzstufen für VERA-3 Deutsch Lesen für Berlin und Brandenburg von 2010 bis 2016, unten Simulation der Verteilung für eine definierte Grundgesamtheit

dargestellt.

Der Schwerpunkt der Verteilung der Itemschwierigkeiten liegt dabei an der gleichen Stelle, wie jener der möglichen Personenfähigkeiten. Welche Bedeutung hat die diskrete Verteilung der möglichen Personenparameter? Üblich kann man erwarten, dass sich die Leistungen der Schülerinnen und Schüler in etwa normal über der Metrik verteilen. Der Mittelwert liegt für die Kompetenzmessung bei den Vergleichsarbeiten erwartungsgemäß im Bereich der zweiten Kompetenzstufe. Schüler*innen, die in der dritten Jahrgangsstufe solche Leistungen zeigen, haben bei durchschnittlichem Lernzuwachs eine gute Chance, den Regelstandard, also die Kompetenzstufe III, bis zum Ende der Jahrgangsstufe 4 zu erreichen. Beide Testhefte de-

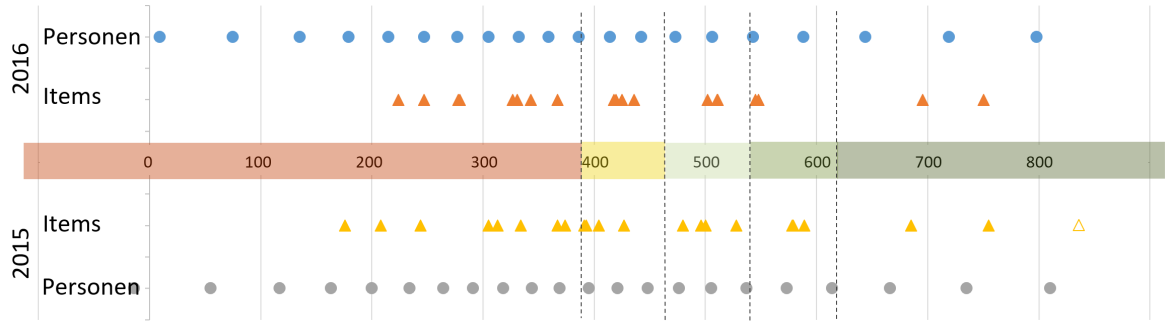


Abbildung 5.7.: Item- und Personenparameter für VERA-3 Deutsch Lesen, 2015 und 2016, Metrik der Bildungsstandards mit 5 Kompetenzstufen

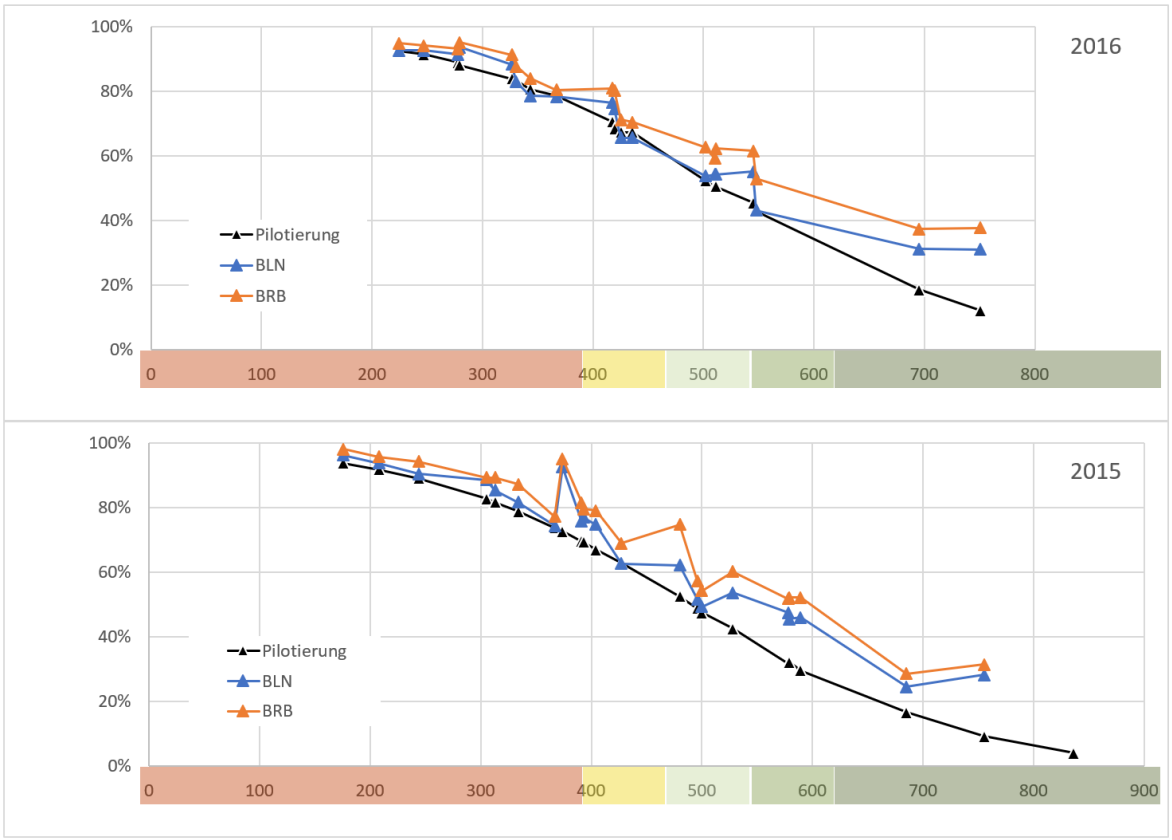


Abbildung 5.8.: Lösungshäufigkeiten für VERA-3 Deutsch Lesen, 2015 und 2016, für Berlin, Brandenburg und aus der Pilotierung

cken mit dem Großteil ihrer Aufgaben genau den Bereich ab, in dem die Leistungen der Schüler*innen zu erwarten sind. Einige wenige Aufgaben sind aber auch deutlich leichter und andere auch deutlich schwerer. Jeweils 8 Items sind im Bereich der Kompetenzstufe I verortet und je 4 in der zweiten. Mit 4 und 3 Items hat das Testheft aus 2015 jeweils ein Item mehr in den Stufen III und IV, als das 2016'er. In der Kompetenzstufe V haben beide Testhefte je 2 Aufgaben.

Nach dem Blick auf die Struktur der Schwierigkeitsverteilung der Items und der daraus resultierenden Verteilung möglicher Personenparameter, werden in der Graphik 5.8 die Lösungshäufigkeiten der Schülerinnen und Schüler Berlins und Brandenburgs für alle eingesetzten Items den Lösungshäufigkeiten der Pilotierung gegenübergestellt. Die Aufgaben sind in der Reihenfolge ihrer Schwierigkeit durch eine Linie verbunden. Deutlich wird sofort, dass bis auf wenige Ausnahmen die Lösungshäufigkeiten beim Einsatz der Testhefte im Rahmen der Vollerhebung bei VERA gleich sind oder in der Mehrzahl über denen aus der Pilotierung liegen. Insbesondere die schweren Aufgaben weisen durchgehend bessere Lösungshäufigkeiten

auf. Das kann ein Hinweis darauf sein, dass die Lehrkräfte ihren Schüler*innen bei schwereren Aufgaben mehr Unterstützung zukommen ließen, als dies die Durchführungsbestimmungen einräumen bzw. dass sie bei der Bewertung hier großzügiger verfuhren. Vermutlich wollen sie dabei insbesondere bei aus ihrer Sicht unangemessen schwierigen Aufgaben einen Ausgleich zur „Strenge des Tests“ schaffen. Allerdings fallen hier keine deutlichen Unterschiede zwischen den zwei Jahren auf. 2015 gibt es im mittleren Bereich zwei Items mit einer Abweichung, die mit mehr als 20% überproportional scheint, die aber für beide Länder ähnlich ausfällt.

Die Metrik, mit der die tatsächliche Leistung gemessen wird, stellt eine stetige Variable dar. Jeder Punkt auf der Skala repräsentiert einen möglichen Fähigkeitswert. Mit einem Messinstrument kann man die Fähigkeit einer Person auf dieser Skala messen. Das bedeutet, dass ein diskreter Wert aus dieser Skala als gemessener Wert eine Fähigkeit repräsentiert. Messinstrumente haben eine begrenzte Genauigkeit und sie zerlegen damit die stetige Skala in eine diskrete. Im Allgemeinen stellt diese Diskretisierung kein Problem dar. Sind die Abstände zwischen zwei diskreten Messpunkten zu groß, wählt man ein Instrument mit angemessen größerer Präzision.

Bei der Messung der Länge eines Tisches mit einem üblichen Maßband, haben die Punkte, an denen der Messwert abgelesen werden kann, einen Abstand von einem Millimeter. Messwerte, die nicht genau auf eine Markierung fallen, werden der nächstliegenden Markierung zugeordnet. Zwischen der realen Länge und der gemessenen Länge liegt bei sonst korrektem Einsatz des Messinstruments maximal ein halber Millimeter Abstand. Alle einem gemessenen Wert zugeordneten Längen kommen aus einem Intervall realer Werte, welches im Beispiel einen Millimeter breit symmetrisch um den gemessenen Wert liegen. Durch die Messung wird aus der stetigen Verteilung von Messwerten also eine diskrete Verteilung.

Auch bei der Skalierung nach Rasch ergeben sich mit einem konkreten Set an Aufgaben nur endlich viele Punkte, an denen die Leistung von Schüler*innen gemessen werden können. Das sind genau die möglichen Personenparameter aus der Abbildung 5.7. Anders als bei einem Maßband sind die möglichen Messwerte allerdings nicht gleichabständig. Deshalb misst das Instrument an unterschiedlichen Stellen der Skala unterschiedlich genau. Wo es genauer misst und wie genau, also an welchen Stellen die möglichen Messwerte liegen, ist ein Resultat der Auswahl der Items (näheres zum Zusammenhang beider Verteilungen ist in Abschnitt 4.1.4 erörtert). Die stetige Verteilung von Fähigkeitswerten wird durch die Messung diskretisiert, wobei unterstellt wird, dass ein Messwert in unmittelbarer Nähe des realen Wertes liegt[14].

[14]Im Abschnitt 4.1.2 wurde zwar gezeigt, dass diese Vorstellung in weiten Bereichen der Skala tatsächlich nicht zutreffend ist. Für die folgenden Erörterungen sind diese Abweichungen aber nicht relevant.

Auch wenn dies praktische Grenzen hat, führt eine Erweiterung des Tests durch zusätzliche Aufgaben mit entsprechenden Schwierigkeitsparametern, zu zusätzlichen Personenparametern, die tendenziell dichter liegen und damit die Genauigkeit verbessern.

Durch die Messung von Fähigkeiten einer Personengruppe (wie alle Schüler*innen einer Klasse oder eines Landes), entsteht durch die der Messung innewohnenden Diskretisierung eine Verteilung der Fähigkeiten auf der Metrik. Dabei werden Personen mit Fähigkeiten in je begrenzten Intervallen zusammengefasst. Diese diskrete Verteilung stellt ein Abbild der realen Verteilung der Leistungen dar. Geschieht diese Abbildung durch zwei Messungen mit verschiedenen Instrumenten und damit verschiedenen möglichen diskreten Messpunkten, werden Vergleiche beider Verteilungen insbesondere dann schwierig, wenn die zwei Diskretisierungen unterschiedlich präzise sind.

Betrachtet man nun die zwei Testinstrumente für das Leseverstehen aus VERA-3 der Jahre 2015 und 2016, ist die Verteilung der Messpunkte bzw. der möglichen Personenparametern dabei ähnlich zu jener der Itemschwierigkeiten. Bei beiden Testheften liegen jeweils 10 Personenparameter im Bereich der Kompetenzstufe I. Auf den Stufen II und III hat das Testheft aus dem Jahr 2015 jeweils 3 und das Testheft aus 2016 nur jeweils 2 Personenparameter. Die Stufen IV und V sind mit 2 und 3 Personenparametern in beiden Testheften wieder gleich besetzt. Solche Unterschiede lassen sich nicht vermeiden, wenn die Testhefte aus unterschiedlich vielen Aufgaben zusammengesetzt sind. Aber selbst wenn die Zahl gleich ist, können die Lagen der Personenparameter in Bezug auf die Kompetenzstufen mehr oder weniger stark differieren.

In der Tabelle 5.2 werden die Charakteristiken der zwei Testhefte (links 2015, rechts 2016) gegenübergestellt, verbunden mit der Auswertung in Berlin und Brandenburg. Zu jedem Testheft werden für die Anzahl korrekt *gelöst*er Aufgaben die *BiSta*-Werte des zugeordneten Personenparameters dargestellt und die entsprechende Kompetenzstufe zugeordnet. Daneben stehen jeweils die Anteile von Schülerinnen und Schülern aus Berlin und Brandenburg, denen dieser Parameter zugeordnet wurde. Diese Verteilungen sind auch in der Graphik 5.9 abgebildet. In der Mitte der Tabelle 5.2 sind zum einen für jedes der zwei Länder die Differenzen der Anteile von 2016 und 2015 dargestellt. In den Kompetenzbereichen I, IV und V, wo die Zahl der in einen Kompetenzbereich fallenden Personenparameter für beide Testhefte identisch ist, sind diese Personenparameter paarweise gegenübergestellt. Für die Kompetenzstufen II und III, in denen das Testheft aus 2015 jeweils einen Personenparameter mehr aufweist, wurden die Anteile innerhalb der Kompetenzstufe für die Differenz summiert. Für eine Bewertung ist

Tabelle 5.2.: Verteilung der Schüler*innen auf die Personenparameter für VERA-3 Deutsch Lesen, 2015 und 2016

Deutsch Lesen 2015					Differenz			Deutsch Lesen 2016				
gel.[a]	Pkt.[b]	KST	Bln	Brb	Bln	Pkt.[b]	Brb	Brb	Bln	KST	Pkt.[b]	gel.
0	-14	1	0.2	0.0	0.1	23	0.0	0.1	0.3	1	9	0
1	55	1	0.2	0.0	0.0	20	0.1	0.1	0.2	1	75	1
2	117	1	0.4	0.2	0.1	18	0.2	0.4	0.5	1	135	2
3	163	1	0.6	0.2	0.2	16	0.3	0.5	0.8	1	179	3
4	200	1	1.1	0.5	0.1	15	0.3	0.7	1.2	1	215	4
5	234	1	1.5	0.8	0.4	13	0.5	1.3	1.9	1	247	5
6	264	1	2.1	1.1	0.7	13	0.6	1.6	2.8	1	277	6
7	291	1	2.7	1.8	1.1	14	0.7	2.3	3.8	1	305	7
8	318	1	3.3	2.1	1.3	14	0.8	2.9	4.6	1	332	8
9	344	1	4.2	3.0	1.0	15	0.9	3.9	5.2	1	359	9
10	369	1	4.9	3.6	1.5	17	1.2	4.8	6.4	1	386	10
11	395	2	5.9	4.6				5.5	6.5	2	414	11
12	421	2	6.3	5.7	-6.5	7	-4.9	6.7	6.9	2	442	12
13	448	2	7.6	6.8								
14	476	3	7.7	9.1				7.8	7.5	3	473	13
15	505	3	8.7	10.1	-10.0	-16	-13.0	8.7	7.8	3	506	14
16	537	3	8.9	10.4								
17	573	4	9.1	11.1	-0.3	-30	-0.7	10.4	8.8	4	543	15
18	614	4	9.2	10.8	0.1	-26	0.9	11.7	9.3	4	588	16
19	666	5	7.6	8.9	2.7	-22	3.5	12.4	10.3	5	644	17
20	735	5	5.2	6.2	4.2	-16	5.2	11.5	9.4	5	719	18
21	810	5	2.4	3.0	3.2	-12	3.7	6.8	5.7	5	798	19

[a]Anzahl korrekt gelöster Items.
[b]Personenparameter für diese Anzahl korrekt gelöster Items liegt bei.

die parallele Betrachtung der graphischen Darstellung 5.9 hilfreich.

Lesebeispiel: Für das Testheft Deutsch Lesen 2015 ist zu entnehmen, dass ein*e Schüler*in bei 9 von 21 korrekt gelösten Aufgaben einen BiSta-Wert von 344 zugeordnet wird, was diese Leistung wiederum der Kompetenzstufe I zuordnet. In Berlin waren genau 4,2% der Schüler*innen hier zu finden, in Brandenburg 3,0%. Ein Jahr später wurde bei 9 korrekt gelösten Aufgaben ein BiSta-Wert von 359 erreicht. Das Testheft enthielt da auch nur 19 Aufgaben. In Berlin fanden sich 2016 5,2% der Schüler*innen mit dieser Leistung. In der Mitte der Tabelle ist ausgewiesen, dass die zwei BiSta-Werte, die bei 9 korrekt gelösten Aufgaben erreicht werden, $359 - 344 = 15$ Punkte auseinander liegen und dass in Berlin 2016 genau $5.2 - 4.2 = 1,0\%$ mehr Schüler*innen diesen Wert erreichten[15].

[15]Hier werden immer die Werte aus dem Jahr 2015 von denen aus dem Jahr 2016 abgezogen.

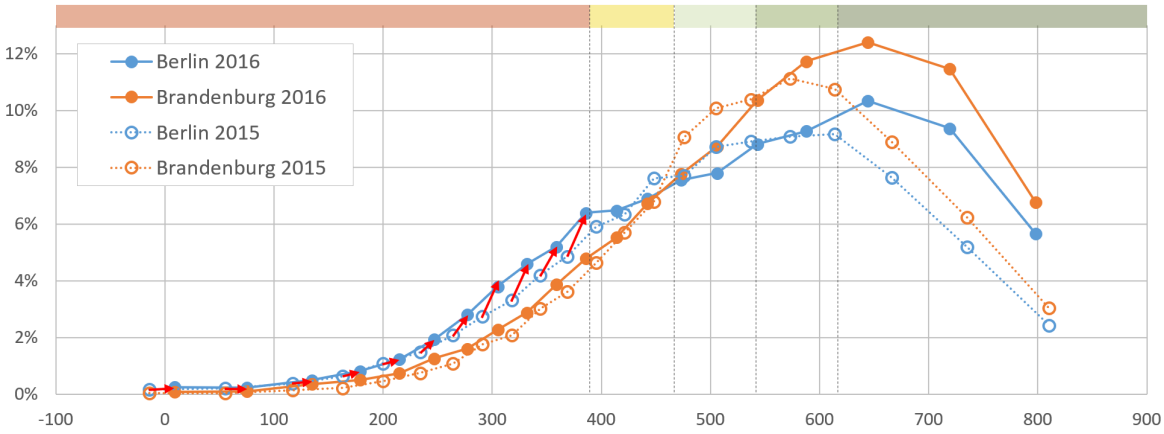


Abbildung 5.9.: Verteilung der Leistungen bei VERA-3 Deutsch Lesen für Berlin und Brandenburg, 2015 und 2016

Die leicht engere Verteilung der Itemschwierigkeiten 2016 führt auch zu einer äquivalent weniger streuenden Verteilung der möglichen Personenparameter (vergleiche auch Abbildung 5.7). So erstrecken sich die Personenparameter des Testhefts von 2015 noch von -14 bis 810 BiSta-Punkten und 2016 über einen um 35 Punkte kleineren Bereich, wobei sich die Mittelwerte der Verteilung der möglichen Personenparameter mit 387 (2015) und 382 (2016) kaum unterscheiden. Was bedeutet dies für die Messung? Geht man davon aus, dass die reale Verteilung der Personenfähigkeiten in etwa einer Normalverteilung entspricht, die mit der Messung an genau den möglichen Personenparametern als Messstellen gemessen wird, dann ist die Lage dieser möglichen Personenparameter als Messstellen von essenzieller Bedeutung für die tatsächlich gemessene Höhe. Im Folgenden werden die Auswirkungen für die Kompetenzstufen einzeln diskutiert.

Die ersten 11 Personenparameter messen für beide Testhefte allein die Personenverteilung im Bereich der Kompetenzstufe I. Um die Kompetenzstufe II zu erreichen, müssen demnach in jedem Fall mehr als 50% der Aufgaben korrekt bearbeitet werden. Im Testheft von 2016 liegen diese 11 Messstellen für die Kompetenzstufe I sämtlich etwas weiter rechts. In der Tabelle 5.2 ist dieser Versatz in der Mitte (Spalten *Differenz BLN* bzw. *BRB*) als itemweise positiver Wert abzulesen. Alle möglichen Personenparameter, also alle Messstellen, im Bereich der Kompetenzstufe I liegen 2016 zwischen 13 und 23 BiSta-Punkten über den paarweise vergleichbaren Messstellen von 2015 und dies, obwohl die Verteilung der Messstellen 2016 im Mittel 5 Punkte unter denen von 2015 liegen. Dass bei unterstellt näherungsweise normalverteilter Fähigkeit Messstellen in der aufsteigenden Flanke einer Verteilung rechts höhere Werte ergeben, ist erwartungsgemäß und hier auch zutreffend. Alle Messwerte ergeben

gleiche oder höhere Anteile, die sich über alle Messstellen innerhalb der Kompetenzstufe I zu dem in der Kompetenzstufenverteilung der Abbildung 5.6 summieren (für Berlin 6,6%, für Brandenburg 5,2%). Der Anstieg des Anteils von Schüler*innen, die der Kompetenzstufe I zugewiesen werden, kann also mindestens teilweise dadurch erklärt werden, dass alle Messstellen weiter rechts liegen. In der Tabelle 5.2 wird dies für das Land Berlin durch die kleinen roten Pfeile angedeutet. Für Brandenburg finden sich ähnliche Verhältnisse.

Für die zweite und dritte Kompetenzstufe gibt es einen weiteren Effekt, der wieder auf die testheftspezifische Diskretisierung der stetigen Metrik der Bildungsstandards zurückgeführt werden muss. Die Anzahl möglicher Personenparameter liegt für das Testheft aus 2015 in beiden Kompetenzstufen jeweils bei 3 und 2016 nur bei 2. So werden in Berlin an den drei Messstellen in der Kompetenzstufe II 5,9%, 6,3% und 7,6%, zusammen also 19,8% gemessen und ein Jahr später an den nur zwei Messstellen mit 6,5% und 6,9% nur 13,4%. Der Anteil ist allein schon deshalb kleiner, weil sich die Zahl der Messstellen reduziert hat. Ähnlich ist dies in der Kompetenzstufe III. Zudem liegt die letzte Messstelle 2015 mit 537[16] nur 3 BiSta-Punkte unter der Kompetenzstufengrenze zur Stufe IV. Umgekehrt liegt die erste Messstelle der Kompetenzstufe IV des 2016er Testhefts mit 543 nur 3 Punkte über der Grenze. Wenn die 2015'er Messstelle 3 Punkte höher bzw. 2016'er 3 Punkte niedriger gelegen hätte, würden die zwei Kompetenzstufen in der Verteilung fast 10% ihrer Anteile tauschen. Bei der Lage von Messstellen ist für die zugeordneten Anteile immer zu betrachten, ob diese Messstellen in einer aufsteigenden oder abfallenden Flanke der Fähigkeitsverteilung liegen oder um dem Mittelwert herum.

Für den oberen Bereich der Kompetenzstufe V findet man das Äquivalent zur Kompetenzstufe I. 2016 liegen die Messstellen an der hier abfallenden Flanke der Verteilung weiter links und vereinen damit jeweils etwas mehr Schüler*innen auf sich. Für Berlin summieren sich diese Differenzen auf 10,1%. Würde im Testheft aus dem Jahr 2015 der letzte Personenparameter der Kompetenzstufe IV von 614 Punkten nur einen einzigen BiSta-Punkt nach oben rutschen[17], würden die als starker Zuwachs wahrgenommenen 10,1% auf ca. 1% abschmelzen.

Um deutlich zu machen, wie groß Schwankungen der Verteilung von Kompetenzstufen sein können, ohne dass dem eine real verschiedene Leistungsverteilung zu Grunde liegt, wurden wieder mit den unterschiedlichen Items-Sets der Jahre 2010 bis 2016 Simulationen vorgenommen. Dazu wurden 25.000 simulierte Schüler*innen mit normalverteilten Fähigkeiten

[16] Die Positionen der Messstellen sind der Tabelle 5.2 zu entnehmen.
[17] Tatsächlich liegt der Personenparameter kaum mehr als einen halben Punkt unter der Kompetenzstufengrenze

die unterschiedlichen Testhefte „vorgelegt“ und ein streng Rasch-konformes Antwortverhalten simuliert. Die Ergebnisse wurden dann jeweils als Verteilung der Kompetenzstufen den Ausgangsdaten gegenübergestellt (siehe Abbildung 5.6 unten). Alle Schwankungen von Stufenanteilen in diesem Bild sind ausschließlich verfahrensbedingt, die zugrundeliegenden Leistungen sind identisch und die Beantwortung frei von jedweden störenden Einflüssen simuliert. Veränderungen von Kompetenzstufenverteilungen wie sie hier zu sehen sind, erlauben demnach keinerlei Interpretation. So ist beispielsweise die nahezu Halbierung des Anteils von Schüler*innen in der Kompetenzstufe I von 2012 zu 2013 keinen tatsächlichen Leistungsunterschieden zuzuschreiben. Dass der Aufwuchs der Anteile von Kompetenzstufe V zwischen den Jahren 2015 und 2016 hier weniger stark auffällt, liegt an der gegenüber der Realität geringeren Besetzung der Kompetenzstufe durch die Simulation.

Unterschiede zwischen zwei Kompetenzstufenverteilungen können verschiedene Gründe haben, von denen tatsächlich unterschiedlich verteilte Leistungen nur einer sind. Es konnte gezeigt werden, dass die Lage der Personenparameter als Messstellen der Verteilung einen enormen Einfluss auf die Kompetenzstufenverteilung hat. So kann auch eine Veränderungen von 10% bei der Besetzung einer Stufe in einigen Fällen nur auf von der Leistungsverteilung unabhängige Effekte zurückgeführt werden. Eine genaue Betrachtung von Kompetenzstufendifferenzen bei einer Messungen mit unterschiedlichen Testheften ist in jedem Fall notwendig. Dies gilt, wie gezeigt werden konnte, sowohl für die Zuordnung einer Kompetenzstufe zur Leistung einer Person, aber auch für beliebige Aggregationen.

5.4.3. Ein Jahr Schule ohne Gewinn für Gymnasiast*innen

Wie im Abschnitt 5.4.1 für Mathematik wird auch hier eine identische Kohorte von Schüler*innen aus der Sekundarstufe einmal auf der Basis von VERA-Daten und ein Jahr später mit Daten aus dem Bildungstrend untersucht und die Entwicklung für die Domäne Deutsch Lesen auf Plausibilität geprüft.

Im VERA-Durchgang 2014 in der achten Jahrgangsstufe Berlins verteilen sich die Leistungen der Schülerinnen und Schüler für die Domäne Deutsch Lesen entsprechend der Spalten *Vergleichsarbeiten 2014 Sek I* und *Gy* der Tabelle 5.3 auf die Kompetenzstufen. Die Spalte *Sek I* fasst dabei die Ergebnisse aller Schülerinnen und Schüler der Sekundarstufe I zusammen, unabhängig von der besuchten Schulform und der damit zugeteilten Testheftvarianten. Die Ergebnisse nur der Schüler*innen an Gymnasien wurde in der Spalte *Gy* dargestellt, deren Ergebnisse erwartungsgemäß deutlich besser ausfallen. Da die zwei Testheftversionen

Tabelle 5.3.: Ergebnisse für Deutsch Lesen bei VERA-8 2014 und beim Bildungstrend 2015

KST	Vergleichsarbeiten 2014				Bildungstrend 2015	
			Stichprobe			
	Sek I	Gy	Sek I	Gy	Sek I	Gy
I	17.1	3.4	15.1	3.4	29.3	7,0
II	22.9	15.1	22.6	15.2	26.9	22.3
III	27.5	33.5	28.5	33.8	25.6	35.6
IV	21.8	32.1	22.7	32.2	14.4	27.1
V	10.7	15.8	11.1	15.4	3.8	8.0
Mw[a]		554		554	483	555
Se	0.79	0.90	1.02	1.17	4.7	6.0
Sd	111	87	110	86	110	81

[a]Die für den Vergleich nicht notwendigen Mittelwerte der gesamten Population werden hier nicht berichtet.

durch die Verankerung auf einer Metrik verortet sind, lassen sich die Ergebnisse direkt vergleichen bzw. zusammenfassen. Überdies sind wieder etwa ein Drittel der Aufgaben beider Testhefte eines Jahres identisch (siehe auch Abschnitt 3.5). Diese Ergebnisse werden im unteren Teil der Tabelle um den Mittelwert, dessen Standardfehler und die Standardabweichung der BiSta-Metrik ergänzt.

Die zwei rechten Spalten der Tabelle 5.3 weisen die Ergebnisse der identischen Kohorte ein Jahr später beim Bildungstrend aus (zu den Messzeitpunkten siehe Tabelle A.7 im Anhang). Hier wurde mit Hilfe einer Stichprobe ebenso die Kompetenz Deutsch Lesen überprüft[18].

Bevor die Ergebnisse der Vergleichsarbeiten mit denen des Bildungstrends ein Jahr später verglichen werden, sind wieder die unterschiedlichen Zielpopulationen zu berücksichtigen. Der Bildungstrend bildet in der Stichprobe sämtliche Schülerinnen und Schüler Berlins ab, die eine neunte Klasse besuchten. Hier werden gleichermaßen öffentliche und private Schulen einbezogen wie auch Schüler*innen aus Förderschulen. Bei Letzteren sollten insbesondere jene berücksichtigt werden, „bei denen davon auszugehen ist, dass die jeweiligen Schülerinnen und Schüler grundsätzlich in der Lage sind, den Test zu bearbeiten."(Stanat et al., 2016, S. 104). Gemeint sind damit, wie weiter ausgeführt wird, Schüler*innen der Förderschwerpunkte *Lernen*, *Sprache* sowie *soziale und emotionale Entwicklung*[19]. Unberücksichtigt bleiben demnach in der Mehrheit zielgleich unterrichtete Schüler*innen der Förderschwerpunkte *Hören*, *Sehen* und *körperliche und motorische Entwicklung*, für deren Teilnahme eine Adaption der

[18]Die Ergebnisse des Bildungstrends für das Land Berlin sind im Anhang A.8 vollständig wiedergegeben.

[19]Schüler*innen dieser Förderschwerpunkte werden vielfach, wenn auch nicht in jedem Fall zieldifferent unterrichtet. Für diese gelten die an die Bildungsstandards angelehnten Rahmenlehrpläne der Länder im Allgemeinen nicht.

Testmaterialien notwendig gewesen wäre.

Tatsächlich bestand die Stichprobe aus 4 (0) Förderschulen, 55 (5) Gymnasien und 65 (3) nicht gymnasialen Schulen[20] (Stanat et al., 2016, S.109). Bei den Vergleichsarbeiten waren hingegen nur zielgleich unterrichtete Schülerinnen und Schüler zur Teilnahme verpflichtet. Schülerinnen und Schüler mit zieldifferenten Rahmenlehrplänen, insbesondere solche an Förderschulen, können freiwillig teilnehmen, ihre Ergebnisse gehen aber nicht in die Berichterstattung ein. Dies führt dazu, dass sich die Teilnahme von Schüler*innen mit verschiedenen Förderbedarfen zwischen den Vergleichsarbeiten und dem Bildungstrend grundlegend unterscheidet. Da die von den Bildungsstandards abgeleiteten Rahmenpläne für zielgleich unterrichtete Schülerinnen und Schüler gelten, werden die Testmaterialien für die Vergleichsarbeiten derart adaptiert, dass möglichst alle diese Schüler*innen einbezogen werden können. Die Adaptionen genügen den im Bildungstrend notwendigen psychometrischen Anforderungen nicht bzw. wird dies nicht überprüft. Festzuhalten bleibt, dass der Einbezug von Schülerinnen und Schülern mit Förderbedarfen bei der Durchführung des Bildungstrends anders geregelt ist, als bei den Vergleichsarbeiten, wo die Partizipation im Vordergrund des Bemühens steht. Die Ergebnisse der Vergleichsarbeiten beziehen zudem nur öffentliche Schulen ein. Nur diese sind zur Teilnahme verpflichtet[21]. So finden sich von den 124 in den Bildungstrend 2015 einbezogenen Schulen lediglich 111 unter den 222 VERA-Schulen des Jahres 2014, darunter keine Förderschule, 50 Gymnasien und 61 Schulen nicht gymnasialer Schulformen.

Eine Gegenüberstellung der Ergebnisse der Vergleichsarbeiten 2014 mit denen des Bildungstrends 2015 sollte wegen der beschriebenen Unterschiede nicht über-interpretiert werden. Für die Teilpopulation der Gymnasiast*innen gilt dies weniger. Förderschulen sind hierbei ausgeschlossen, nur vereinzelt sind integrierte Schüler*innen mit Förderbedarfen einbezogen. Bis auf die 5 Schulen in privater Trägerschaft können alle 50 öffentlichen Gymnasien der Stichprobe des Bildungstrends auch in den Daten der Vergleichsarbeiten gefunden werden. Dies wird sicherlich dadurch begünstigt, dass für die Stichprobenziehung im Bildungstrend die Schulstatistik des vorhergehenden Jahres verwendet wird, also genau die Daten aus dem Jahr der VERA-Durchführung. Berechnet man für die 50 auch im Bildungstrend beteiligten Gymnasien die Kompetenzstufenverteilung wie die Maße der zentralen Tendenz auf der Metrik der Bildungsstandards, so wie das auch beim Bildungstrend geschieht, ergeben sich die Werte der Spalte *Vergleichsarbeiten 2014, Stichprobe, Gy* der Tabelle 5.3. Die Kongruenz der Werte mit den Ergebnissen aller an VERA teilnehmenden Schülerinnen und Schülern aus Gym-

[20] In Klammern sind jeweils die darin enthaltenen Schulen in privater Trägerschaft ausgewiesen.
[21] Über die Auswahl freiwillig teilnehmender privater Schulen gibt es keine Informationen.

nasien zeugt von einer hervorragenden Stichprobe, die von ca. 11.000 Schüler*innen aus 90 Gymnasien je eine Klasse aus 50 ausgewählten Gymnasien zieht.

Während sich die Verteilung der Leistungen auf die Kompetenzstufen für die Gymnasiast*innen zwischen den Vergleichsarbeiten in der achten und dem Bildungstrend in der neunten Jahrgangsstufe unterscheiden, sind die Mittelwerte beider Erhebungen aber nahezu identisch. Zwischen beiden Messungen liegt allerdings mehr als ein Schuljahr[22]. Wo zum Ende der Sekundarstufe für ein Schuljahr ein Zuwachs von 15 bis 20 Punkten erwartet wird (Stanat et al., 2016, S. 335), findet sich quasi keinerlei Veränderung. Diese Ergebnisse und auch die aller Schüler*innen der Sekundarstufe I aus Schulen der Bildungstrend-Stichprobe lassen bei aller Vorsicht wegen der Unterschiedlichkeit der Zielpopulationen vermuten, dass auch hier die inzwischen bekannten Unplausibilitäten bestehen. Allerdings lohnt sich die genaue Betrachtung der Zahlen kaum, denn auch für die Domäne Deutsch Lesen bei VERA-8 finden sich Variabilitäten über der Zeit, wie diese schon für Mathematik gezeigt wurden (siehe Abschnitt 5.4.1). Problematisch ist also weniger die gerade hier konkretisierte unplausible Differenz als die Vermutung, dass es in einem anderen Jahr eine zufällig andere Differenz gewesen wäre.

5.5. Untersuchung der Stabilität bei mehrfachem Einsatz verschiedener VERA-Tests

Der vorhergehende Abschnitt hat beispielhaft einige unplausible Ergebnisse bei der Messung von Leistungsständen im Rahmen von Vergleichsarbeiten gezeigt. In den folgenden zwei Teilabschnitten sollen zwei im Rahmen der Vergleichsarbeiten durchgeführte Untersuchungen des ISQ mit dem Ziel dargestellt werden, die Stabilität von VERA-Instrumenten auf der Basis deren mehrfachen Einsatzes zu bewerten. Beide Studien sind ursprünglich nicht zu diesem Zweck initiiert worden. Allerdings wurde wegen Zweifel an der Stabilität der VERA-Instrumente für die Studie VERAMSA die Auswahl der verwendeten VERA-Instrumente im dritten Testzeitpunkt variiert, woraus sich die Möglichkeit der hier ausgeführten Untersuchung ergab. In der zweiten Studie, der Entwicklung von Orthographiekompetenz, wird das identische Instrument ein zweites Mal eingesetzt, um Leistungsentwicklung abzubilden (Vettorazzi & Harych, 2019). Wird hierbei gezeigt, dass dieses klassische Untersuchungsdesign funktioniert, kann beim wiederholten Einsatz eines Instruments von einer hinreichenden Stabilität ausgegangen werden.

[22] Die Testungen zu VERA-8 fanden am 21. März 2014 statt. Der Testzeitraum für die Erhebungen zum Bildungstrend erstreckte sich für Berliner Schulen vom 11. Mai bis zum 17. Juni 2015.

5.5.1. Untersuchung der Stabilität beim Einsatz verschiedener, aber miteinander verlinkter VERA-Instrumenten

Fragestellung

Konkret soll hier der Frage nachgegangen werden, wie stabil die Verlinkung von VERA-Testinstrumenten ist, wenn diese unter restriktiven Durchführungsbedingungen eingesetzt werden. Auch wenn die Datenbasis nicht primär zum Zwecke der Untersuchung der Stabilität von VERA-Instrumenten konzipiert wurde, lassen sich mit einer Re-Analyse der Daten aus der VERAMSA-Studie (Graf et al., 2016) unter Hinzuziehung bisher ungenutzter Daten zumindest Verdachtsmomente ausräumen oder erhärten.

VERAMSA – eine Panelstudie mit Instrumenten aus VERA-8

Um die Prognosegüte der Vergleichsarbeiten für die Ergebnisse der Prüfungen zum Mittleren Schulabschluss Ende der zehnten Klasse zu untersuchen, startete das ISQ 2011 eine Panelstudie. Dazu wurden die Vergleichsarbeiten für ausgewählte Stichprobenschulen anders als üblich durch Testleitungen administriert und die Testhefte zentral durch geschulte Kodiererinnen und Kodierer bewertet. Etwas mehr als zwei Jahre später wurden zum Ende der Jahrgangsstufe 10 die Jahrgangs- und Prüfungsergebnisse erfasst und schüler*innenweise den Ergebnissen der ersten Erhebung zugeordnet. Theoretischer Hintergrund, Fragestellung und Ergebnisse finden sich bei (Graf et al., 2016) wie auch der Hinweis, dass bei den gleichen Schüler*innen zusätzlich auch in der Klasse 9 und kurz vor der Prüfung zum Mittleren Schulabschluss in Klasse 10 Leistungstests mit VERA-Testheften des Faches Mathematik durchgeführt wurden. Verwiesen wird zudem auf Graf et al. (2013) für eine an die Erhebungen zu VERAMSA angeschlossene Untersuchung. Mit den folgenden Analysen, werden alle VERA-Daten dieser längsschnittlich verbundenen Erhebungen betrachtet und damit auch zwei bisher noch nicht ausgewertete Datensätze des zweiten und dritten Messzeitpunktes.

Methode

Stichprobe und Testzeitpunkte Für das zentrale Anliegen der Studie war ursprünglich neben der Erhebung des Leistungsstandes mit VERA-Instrumenten zum Anfang des Untersuchungszeitraumes nur die Erfassung der Prüfungsergebnisse notwendig. Zu beiden Testzeitpunkten wurden alle vorliegenden Test- und Prüfungsergebnisse erhoben, im Falle der Vergleichsarbeiten Mathematik, Deutsch-Lesen und für Englisch als erste Fremdsprache Lese- und Hörverstehen. Weil bei den Vergleichsarbeiten nur für Berlin die Durchführung für alle

Fächer und Testbereiche verpflichtend war, fokussierte die Studie Schulen aus Berlin. Um die Schulform stabil zu halten, wurden 12 Gymnasien mit Hilfe einer stratifizierten Stichprobe ausgewählt. Dazu wurden die MSA-Ergebnisse des Vorjahres der Grundgesamtheit aller öffentlichen Berliner Gymnasien in vier leistungsbezogene Quartile unterteilt und aus jedem Quartil drei Gymnasien zufällig ausgewählt. Aus diesen 12 Gymnasien wurden dann alle 48 achten Klassen mit sämtlichen Schülerinnen und Schülern (N=1306) in die Untersuchung einbezogen. Wie oben angedeutet wurden auf Grund weiterer Fragestellungen in diesem Panel auch in der neunten (N=1200) und in der zehnten Jahrgangsstufe (N=1131) zusätzliche Erhebungen mit verschiedenen Mathematik-Tests aus Vergleichsarbeiten durchgeführt (Graf et al., 2016). Auch diese erfolgten testleitungsbasiert und wurden durch geschulte Kodierer*innen und Kodierer bewertet. Da in der vorliegenden Untersuchung Zugewinne zwischen den einzelnen Messzeitpunkten und diese allein auf der Basis der VERA-Instrumente betrachtet werden sollen, wurden keine Ergebnisse aus der ursprünglichen VERAMSA-Studie verwendet.

Auswahl der Testhefte Für die hier anzustellenden Analysen sind die VERA-Datensätze aller drei Messzeitpunkte und die Konstellationen der unterschiedlichen eingesetzten Instrumente essentiell und sollen folgend dezidiert dargestellt werden. Im Jahr 2011 (T_1 - 2011) wurde als Baseline an den ausgewählten Gymnasien mit dem Testheft 2 das für Berlin in diesem Jahr in allen Gymnasien standardmäßig eingesetzte Heft verabreicht. Für die Folgeerhebung ein Jahr später (T_2 - 2012) wählte man aus dem für das Jahr 2012 vom IQB zur Verfügung gestelltem Set mit dem Testheft 3 das schwierigste Heft aus, um Deckeneffekte zu vermeiden, die insbesondere an den Gymnasien zu befürchten waren. Im dritten Jahr (T_3 - 2013) sollte kurz vor der Prüfung zum Mittleren Schulabschluss ein letzter Mathematik-Test mit VERA-Instrumenten durchgeführt werden. Da unklar war, wie sich die damals zwischen den Jahren relativ großen Schwankungen der Testheftschwierigkeiten (vergleiche Abbildung 4.5) auf die Ergebnisse auswirken würden, wurden 2013 drei unterschiedliche Testhefte eingesetzt (siehe Tabelle 5.4).

Aufbau der Testhefte und Stichprobe von Subpopulationen Im Studienzeitraum wurden vom IQB jedes Jahr drei Testhefte unterschiedlicher Schwierigkeit zur Verfügung gestellt. Diese Testhefte werden dabei aus 4 Testblöcken zusammengesetzt, wobei sich je zwei Hefte benachbarter Schwierigkeit einen Testblock teilen. Das IQB sieht das dritte Testheft für Gymnasien vor, welches sich aus den Blöcken 3 und 4 zusammensetzt. Berlin und Brandenburg setzen hingegen weitestgehend das mittelschwere Testheft 2 aus den Blöcken 2 und 3 für

Tabelle 5.4.: Testhefteinsatz von Mathematik-Tests im Rahmen von VERAMSA

VERA-Testhefte im Studienzeitraum										
Jahr		2011			2012			2013		
Block		2	3	4	2	3	4	2	3	4
Items		22	16	16	20	16	18	21	20[a]	22
BiSta		446	587	598	440	535	657	381	545[a]	587
... für Testheft	2	505			482			461		
	3	592			600			567[a]		
Testhefteinsatz in der Studienstichprobe zu den drei Testzeitpunkten										
T_1 - 2011		alle								
T_2 - 2012					alle					
T_3 - 2013		G1			G2			G3		

[a]Das letzte Item des Blocks 3 wurde weder im Testheft 2 noch im 3 verwendet. Mit diesem Item beträgt der Blockmittelwert 554 und der Mittelwert für das zweite Testheft 467 und für das dritte Testheft 571.

Gymnasien ein. In der Tabelle 5.4 sind für die drei betreffenden Jahre jeweils die Blöcke 2, 3 und 4 für die Testhefte 2 und 3 dargestellt. Die Blöcke und Testhefte werden in der Tabelle bezüglich ihrer Schwierigkeit charakterisiert. So wird beschrieben, aus wie vielen Items ein Testblock besteht und wie hoch die mittlere Schwierigkeit in BiSta-Punkten ist. Darunter ist für die zwei Testhefte jeden Jahres die mittlere Schwierigkeit angegeben. Im unteren Teil der Tabelle ist für jeden der drei Messzeitpunkte T_1, T_2 und T_3 angegeben, welche Instrumente im Untersuchungsdesign eingesetzt wurden. Zu den ersten zwei Testzeitpunkten wurde allen Schülerinnen und Schülern der gesamten Stichprobe jeweils das identische Testheft vorgelegt. Für den letzten Zeitpunkt wurde die Stichprobe in drei Teile unterteilt. Der ersten Gruppe (G1) mit etwa der Hälfte der Teilnehmer*innen wurde das Testheft 3 aus dem Jahr 2011 vorgelegt. Die eine Hälfte dieses Testhefts (Block 3) war auch Teil des Testhefts 2, welches den Schülerinnen und Schülern zum Studienbeginn im Jahr 2011 vorgelegt wurde. Eine zweite Gruppe (G2) mit einem Viertel der Studienteilnehmer*innen erhielt das Testheft 3 des Vorjahres, also jenes Testheft, welches ein Jahr zuvor von allen Studienteilnehmer*innen bearbeitet worden ist. Die letzte Gruppe (G3) erhielt das 2013 vom IQB neu zur Verfügung gestellte Testheft 3. Die Gruppenteilung erfolgt zufällig auf Ebene der Schülerinnen und Schüler, die Testheftversionen wurden also innerhalb jeder Klasse gestreut. Die Abbildung 5.10 stellt die Teilnahmen zu den drei Messzeitpunkten dezidiert dar, so dass alle Zusammenstellungen abgeleitet werden können. Auf eine Unterscheidung zwischen zu einzelnen Zeitpunkten neu hinzugekommenen oder ausgeschiedenen Schüler*innen oder solchen, die nur am Testtag nicht anwesend waren wurde hier aus Gründen der Übersichtlichkeit verzichtet. Dass die Aufteilung

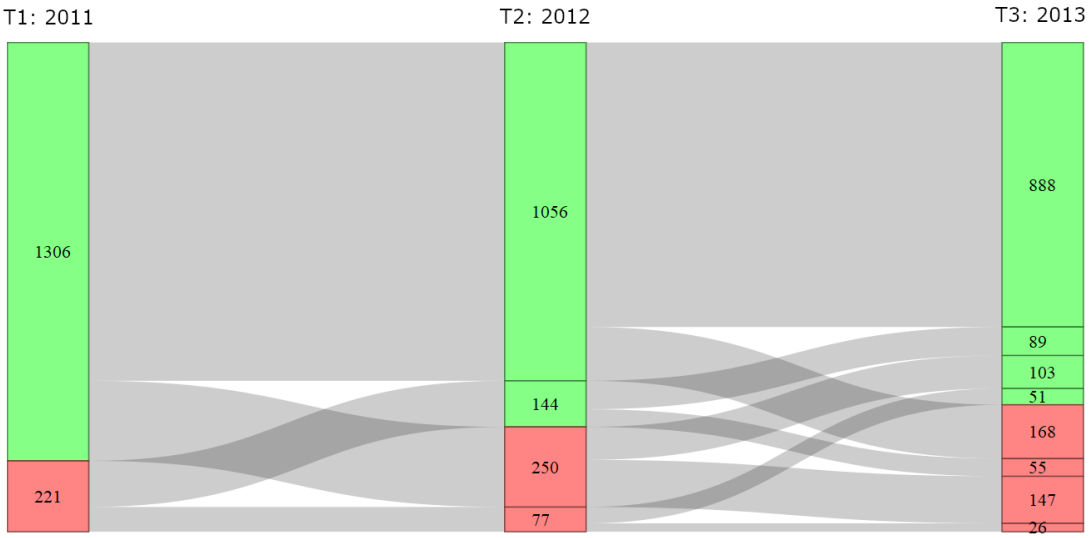


Abbildung 5.10.: Ausfallanalyse über die drei Testzeitpunkte

Tabelle 5.5.: Testheftverteilung zum Zeitpunkt T3 (2013)

	2013	Testheftvarianten 2013		
		G1	G2	G3
	888	444	228	216
	89	42	24	23
	103	51	26	26
	51	24	14	13
Summe	1131	561	292	278
rel		49,6	25,8	24,6

der Testheftvarianten zum letzten Messzeitpunkt eine gelungen zufällige Aufteilung darstellt, erkennt man an der äquivalenten Aufteilung auf die Teilnahmegruppen in der Tabelle 5.5.

Untersuchungsdesigns mit identischem bzw. direkt verlinktem Instrument Aus diesem Design ergeben sich nun unterschiedliche Möglichkeiten der Analyse, die im Folgenden dezidiert dargestellt werden. Hierbei ist in jedem Fall folgende Frage zu beantworten: Welche zwei in welcher Form miteinander verlinkten Instrumente messen Leistungsunterschiede welcher Kohorte zwischen welchen zwei Messzeitpunkten?

Zugewinn von T1 nach T3 Zuerst einmal kann für die Gruppe G1 ein Kompetenzzugewinn über den vollständigen Studienzeitraum von 2 Jahren auf der Basis der 16 identischen Items aus dem Block 3 des VERA-Tests von 2011 ermittelt werden. Aus der Abbildung 5.10 kann man entnehmen, dass von den 1306 zum Testzeitpunkt T1 teilgenommenen Schülerin-

nen und Schülern 1056 im Jahr 2012 und von diesen 888 auch 2013 teilgenommen haben. Von den 2011 verbleibenden 250 Schüler*innen, die 2012 nicht teilgenommen haben, haben 103 aber 2013 Testergebnisse geliefert, So dass insgesamt von 991 Schülerinnen und Schülern Testergebnisse sowohl aus 2011 sowie aus 2013 vorliegen. Von diesen gehören 495 (der Tabelle 5.5 als Summe 444 + 51 zu entnehmen) zur Testgruppe G1, denen 2013 das Testheft 3 aus dem Jahr 2011 vorgelegt worden ist. Für diese Schüler*innen kann der individuelle Leistungszuwachs festgestellt werden. Zwar ist die Basis mit 16 Items, also einem halben VERA-Testheft, eher klein und die Repräsentation der fachlichen Aspekte noch einmal eingeschränkter, als dies schon dem ganzen VERA-Testheft unterstellt werden könnte. Allerdings sollte durch die Nutzung von identischen Items die Messung eines Zuwachses stabil möglich sein. Die anderen Items in den jeweiligen Testheften der zwei Erhebungszeitpunkte T1 und T3 sind mit den 16 Items über die Schwierigkeitsparameter direkt verlinkt. In einer zweiten Analyse des Zuwachses werden nicht nur die identischen 16 Items aus dem Block 3 von 2011 verwendet, sondern alle Items des jeweiligen Testhefts (Testheft 2 aus 2011, verwendet zu T1 2011 sowie Testheft 3 aus 2011, verwendet zu T3 2013). Ein Unterschied festgestellter Zuwächse zwischen der ersten auf 16 Items und dieser zweiten auf zusätzlichen aber unterschiedlichen Items beruhenden Analysen resultiert aus der differenten Repräsentanz des Inhalts durch die Items einmal des halben und dann des ganzen Testheftes. Beide Untersuchungen werden noch einmal wiederholt, wobei dabei alle Schülerinnen und Schüler einbezogen werden, die am Test 2011 (N=1306) bzw. 2013 (N=561) teilgenommen haben, unabhängig davon, welches konkrete Heft bearbeitet wurde. Unterschiede resultierten hierbei zusätzlich aus der differenten Repräsentanz der Schüler*innen, wobei in allen Fällen lediglich Berliner Gymnasiast*innen die Grundgesamtheit darstellen.

Zugewinn von T2 nach T3 Einfacher ist die Analyse des Kompetenzzugewinns zwischen den Klassen 9 und 10 für die Studienteilnehmer*innen, die der Gruppe 2 zugeordnet wurden, denn zu beiden Testzeitpunkten wurde das gesamte Testhefts 3 aus dem Jahr 2012 eingesetzt. Inhaltlich ist mit dem gesamten Testheft eine fachliche Abdeckung gegeben, wie sie bei den Vergleichsarbeiten üblich ist. Da den Testheften grundsätzlich die Standards der Klasse 10 zu Grunde lagen und hier zudem das schwierigste Testheft eingesetzt wurde, sollten Deckeneffekte weitestgehend vermieden werden. Die Stichprobe besteht aus den 252 Schülerinnen und Schülern (wieder als Summe 228+24 aus der Abbildung 5.10 und Tabelle 5.5 zu entnehmen), die zum Zeitpunkt T3 der Gruppe G2 angehören und auch 2012 teilgenommen haben. Auch diese Untersuchung wurde mit allen Schüler*innen wiederholt, die am jeweiligen Testzeit-

punkt das Testheft bearbeitet haben. 2012 waren es 1200 Schüler*innen, 2013 mit 292 nur ein Viertel der Untersuchungsgruppe.

Da in diesen beiden Fällen der Feststellung des Kompetenzzugewinns vollständig oder zumindest teilweise identische Instrumente verwendet werden, wird diese Form der Verlinkung als *direkt* bezeichnet. Demgegenüber ist die Verlinkung von disjunkten VERA-Instrumenten verschiedener Jahre, wie sie im Abschnitt 3.6 (siehe auch Abbildung 3.4) dargestellt wurde *indirekt*, weil zwei VERA-Instrumente erst durch zwei Prozesse über die „Brücken" von mit den Instrumenten selbst disjunkten Ankeritems verlinkt werden.

Untersuchungsdesigns mit unterschiedlichen Instrumenten Alle bisher geplanten Analysen des Zuwachses nutzen identische oder zumindest teilweise identische Instrumente, sind also direkt miteinander verlinkt und daher als sehr stabil einzuschätzen. Diese Art der längsschnittlichen Messung liegt auch beim Bildungstrend vor. Dort werden zwar unterschiedliche Schülerpopulationen untersucht, aber durch die Nutzung der größtenteils identischen Instrumente, sind die Ergebnisse direkt und ohne Einschränkungen vergleichbar. Schlussfolgerungen müssen lediglich eine jeweils korrekte Stichprobenziehung unterstellen. Bei einer Untersuchung der Zuwächse mit verschiedenen Instrumenten, deren Ergebnisse erst durch einen Prozess der Verlinkung auf einer Metrik abgebildet und damit vergleichbar werden, muss der Prozess der Verlinkung selbst einer Prüfung unterzogen werden. Dazu werden die Zugewinne unter Nutzung aller vorliegenden Testergebnisse, also auch mit Ergebnissen aus unterschiedlichen Testheften, ermittelt. Bei funktionierender Verlinkung der Instrumente sollten sich die Ergebnisse der ersten Berechnung erhärten.

Zugewinn von T1 nach T2 und T2 nach T3 Es wird für die Stichprobe der Zugewinn von Jahrgangsstufe 8 nach 9 auf der Basis aller Schülerinnen und Schüler ermittelt, die an beiden Testzeitpunkten T_1 und T_2 teilgenommen haben (N=1056). Bei der zweiten Untersuchung des Zuwachses zwischen Jahrgangsstufe 9 nach 10 bleibt die Aufteilung der Stichprobe zum Zeitpunkt T_3 auf verschiedene Instrumente zunächst unberücksichtigt (N=977). Anschließend wird untersucht, ob sich für die unterschiedlichen Testhefte der drei Gruppen differente Zuwächse ergeben. Hierbei kann der Zuwachs von 8 nach 9 dem von 9 nach 10 gegenübergestellt werden. Dabei geht es nicht um einen Vergleich der Zuwächse selbst, für deren Unterschiedlichkeit sich Gründe finden ließen, sondern um die Unterschiede der Zuwächse zwischen den drei Gruppen. Hier stellt die Messung des Zugewinns von 8 nach 9 quasi eine Basiserhebung dar. Unterschiedlichkeiten zwischen den Daten der drei Gruppen

Tabelle 5.6.: Ergebnisse des Kompetenzzugewinns von T1 (2011) nach T3 (2013) mit identischen bzw. direkt verlinkten Instrumenten

	Instrument[a]	N	Mw	Sd	Se	CI	Diff	CI
T1	Block 3 (2011)	495	535,3	95,9	56,8	[526,8 – 543,8]	63,5	[46,1 – 80,9]
T3	Block 3 (2011)	495	598,8	101,6	57,4	[589,8 – 607,8]		
T1	Heft B (2011)	495	536,1	88,9	38,7	[528,3 – 543,9]	52,1	[35,6 – 68,5]
T3	Heft C (2011)	495	588,2	97,5	40,1	[579,6 – 596,8]		
T1	Block 3 (2011)	1306	525,4	97,7	57,1	[520,1 – 530,7]	67,5	[53,9 – 81,2]
T3	Block 3 (2011)	561	592,9	100,9	57,2	[584,5 – 601,3]		
T1	Heft B (2011)	1306	521,6	90,9	38,4	[516,7 – 526,5]	60,5	[47,4 – 73,6]
T3	Heft C (2011)	561	582,0	98,4	40,0	[573,9 – 590,2]		

Oben: Basis sind Schüler*innen mit Teilnahmen an beiden Testzeitpunkten. Unten: Basis sind alle Schüler*innen mit Teilnahme nur am jeweiligen Testzeitpunkt.
[a]Block: Skalierung und Differenzberechnung auf der Basis der 16 gemeinsamen Items. Heft: Skalierung und Differenzberechnung auf der Basis aller Items des jeweiligen Testhefts, die 16 gemeinsame Items haben.

sind hierbei eben nur auf die Unterschiedlichkeit der Gruppen zurückzuführen, denn die für die Gruppen verwendeten Instrumente unterscheiden sich für den ersten Zeitbereich von T_1 nach T_2 nicht. Bei funktionierender Verlinkung der Instrumente sollte die Streuung zwischen den drei Instrumenten bei der Messung des Zuwachses von Jahrgang 9 nach 10 nur zufällig anders sein. Wegen der kleinen Zahl von Probanden können die Konfidenzintervalle allerdings so groß werden, dass Unterschiede sich ggf. als nicht signifikant erweisen. Da diese Form der Auswertung zum Erhebungszeitpunkt nicht intendiert war, fand auch keine Festlegung einer für diesen Untersuchungszweck angemessenen Stichprobengröße statt.

Ergebnisse

Mit Bezug auf Stanat et al. (2019a, S. 201) wurde der Zugewinn für das Ende der Sekundarstufe I mit etwa 50 Punkten angegeben. Die mit den ersten beiden Analysen zu ermittelnden Zugewinne von Jahrgangsstufe 8 nach 10 und von Jahrgangsstufe 9 nach 10 sollten pro Jahr also in diesem Bereich liegen, vielleicht auch für Berlin und für die hier untersuchte Stichprobe von Gymnasiast*innen angemessen davon abweichen.

Untersuchung des Zuwachses mit identischen bzw. direkt verlinkten Instrumenten. Die Ergebnisse der Untersuchung des Kompetenzzugewinns zwischen 2011 und 2013 sind in der Tabelle 5.6 zusammengefasst. Die Tabelle 5.7 zeigt die Analyse des Zugewinns zwischen T2 und T3.

Zugewinn von T1 nach T3 Die zwei oberen Zeilen beschreiben die Messungen zum Zeitpunkt T1 (erste Zeile) und T3 (zweite Zeile) und mithin in den letzten zwei Spalten den Zugewinn inkl. des Konfidenzintervalls. Die Basis dieser Messung ist die restriktivste. Hier werden nur jene 495 Schüler*innen einbezogen, die an beiden Zeitpunkten das identische Material bearbeitet haben. Auch auf der Ebene der Testinhalte wurden nur jene 16 Aufgaben des Blocks 3 der Testhefte aus dem Jahr 2011 einbezogen, die zu beiden Zeitpunkten von diesen Schüler*innen bearbeitet wurden. Das Ergebnis zeigt einen leicht unter den Erwartungen liegenden Zuwachs von 63,5 Punkten, mit einem erwartbar großen Konfidenzintervall. Das Ergebnis ist aber insofern plausibel, dass sich die Erwartung des Zuwachses auf die Feststellung des Zuwachses einer bundesweiten Normstichprobe von Schüler*innen bezieht und auch hier nur auf den Zuwachs von Klasse 9 nach 10. In der Untersuchung hingegen wurde der Zuwachs lediglich für Schüler*innen aus Berlin und auch hier nur an Gymnasien bestimmt, allerdings von Klasse 8 nach 10. Einschränkend ist zudem, dass die 16 Aufgaben 2011 die zweite Hälfte des Testhefts bildeten und 2013 die erste. Wenn man annimmt, dass die Aufgaben im hinteren Testteil potentiell weniger gut bearbeitet werden, könnte der Zuwachs auch deshalb leicht geringer ausgefallen sein.

In den Zeilen 3 und 4 ist die gleiche Analyse über die jeweils vollständigen Testhefte erfolgt. Die Repräsentation der Inhalte ist damit zwar größer, ggf. aber zwischen den zwei Messungen auch weniger identisch. Die Ergänzung des Blocks 3 im 2011er Test sollte eher leichtere Aufgaben enthalten, für das 2013er Testheft gilt das Gegenteil. Dies könnte die Ergebnisse gut erklären, denn die Messung am Ausgangspunkt zu T1 unterscheidet sich mit einem Punkt nur marginal, während der Zugewinn durch den Abfall der Leistungen zum Zeitpunkt T3 um gute 10 Punkte deutlich ausfällt. Auch diese erweiterte Messung zeigt durchaus plausible Ergebnisse, verbessert die Messung des Zugewinns qualitativ aber nicht.

Im zweiten Block der Ergebnistabelle 5.6 sind die Analysen noch einmal wiederholt worden, allerdings unter Hinzuziehung sämtlicher 1.306 Personen, die an jedem der Zeitpunkte das entsprechende Material bearbeitet haben, also ohne Rücksicht darauf, ob die Person zum jeweils anderen Messzeitpunkt Teil der Studie war. Für die erste Messung (T1) auf der Basis nur des Blocks 3 ergeben sich ca. 10 Punkte weniger. Wie lässt sich dies erklären? Von den 1.306 Personen haben 991 Personen sowohl an T1 wie T3 teilgenommen. Diese Gruppe teilt sich in jene 495 oben schon beschriebenen Personen auf, die zu T3 den gleichen Block 3 bearbeitet haben und jene 496, denen jeweils eines der zwei anderen Testhefte vorgelegt wurde. Diese 406 Personen erreichen zum Ausgangszeitpunkt mit 537,0 fast das gleiche Ergebnis, wie

die erste Gruppe. Das spricht für eine sehr gute Aufteilung der Gruppe. Die verbleibenden ca. 24% der Personen (315) müssen dann aber deutlich schwächer abgeschnitten haben, um die 10 Punkte geringere Performance der Gesamtgruppe zu erklären. Und tatsächlich erreichen diese Schüler*innen nur 470,4 Punkte und damit ca. 65 Punkte weniger, was mehr als einem Schuljahr entspricht oder in etwa dem Zugewinn, den die anderen Schüler*innen in den folgenden zwei Jahren erfahren werden. Von diesen 315 Schüler*innen haben 147 ihre jeweilige Klasse nach dem 8. Jahrgang verlassen und 168 erst nach der Klassenstufe 9. Letztere erreichen denn auch zum Zeitpunkt T1 immerhin noch 510,1 und damit nur 25 Punkte weniger, als jene, die bis zum zehnten Jahrgang im Klassenverband verbleiben. Diese Dropouts sind sehr plausibel. Vermutlich haben diese Schüler*innen aus offensichtlich leistungsbezogenen Gründen entweder eine Klassenstufe wiederholt oder auch das Gymnasium als Schulform verlassen. Damit ist auch klar, warum der Zugewinn in dieser Analyse insgesamt höher ausfällt. Nicht etwa, weil die Schüler*innen mehr hinzugewinnen, sondern, weil die am Anfang schlecht performenden Schüler*innen aus dem weiteren Studienverlauf ausgeschlossen sind, das Ergebnis der Basiserhebung aber drücken.

Die Zahl an Schüler*innen, die zum Zeitpunkt T3 Teil der Untersuchungsgruppe sind, aber nicht an T1, ist mit 66 deutlich kleiner. Ein Teil (N=24) hat erst zum Zeitpunkt T3 die Testgruppe ergänzt, ein zweiter Teil (N=42) ist unmittelbar nach der ersten Testung dazugekommen. Beide Gruppen unterscheiden sich wieder sehr deutlich in den Mittelwerten von den anderen Schüler*innen, die zu T3 einen Mittelwert von fast 600 zeigten. Die zwei Gruppen liegen mit 519,7 und 565,1 deutlich darunter, allerdings verändern diese 66 Schüler*innen mit ihren knapp 12% Anteil an der Testgruppe die Ergebnisse weniger, als dies für die Messungen zum Zeitpunkt T1 der Fall ist. Um welche spezifischen Gruppen von Schüler*innen es sich hierbei handelt, kann nur vermutet werden[23]. Konkrete Daten liegen dazu nicht vor.

In der Summe ergeben sich beim Hinzuziehen von allen möglichen Teilnehmer*innen zu beiden Zeitpunkten schlechtere Werte. Weil sich dies aber stärker zu T1 als zu T3 auswirkt, ergibt sich ein leicht größerer Zuwachs. Die Ergebnisse sind von der Größenordnung her aber ähnlich (siehe Tabelle 5.6, untere Hälfte). Ebenso ergeben sich leichte Veränderungen, wenn der Zuwachs nicht allein auf der Basis der 16 Items ermittelt wird, sondern das jeweils gesamte Testheft hinzugezogen wird. Für die Messung zum Zeitpunkt T1 scheint der Unterschied vernachlässigbar klein. Hier werden die 16 Items durch einen zweiten Block mit tendenziell einfacheren Items ergänzt. Ein leichter Leistungsverlust zeigt sich hingegen für die Messung

[23]Naheliegend ist, dass ein Schuljahr zurückgestellte Schüler*innen zur Wiederholung in die Klasse kommen.

Tabelle 5.7.: Ergebnisse des Kompetenzzugewinns von T2 (2012) nach T3 (2013) mit identischem Instrument

Instrument	N	Mw	Sd	Se	CI	Diff	CI
T2 Heft C (2012)	252	546,8	85,1	40,3	[536,3 – 557,4]	23,8	[1,3 – 46,4]
T3 Heft C (2012)	252	570,7	97,5	41,2	[558,7 – 582,7]		
T2 Heft C (2012)	1200	536,3	91,9	40,7	[531,1 – 541,5]	31,9	[15,4 – 48,3]
T3 Heft C (2012)	292	568,1	98,0	41,2	[556,9 – 579,4]		

Oben: Basis sind Schüler*innen mit Teilnahmen an beiden Testzeitpunkten. Unten: Basis sind alle Schüler*innen mit Teilnahme am jeweiligen Testzeitpunkt.

zum Zeitpunkt T3. Hier wird das Testheft durch den zweiten Block allerdings schwerer. Dies allein sollte allerdings bei Rasch-konformen Randbedingungen keine schlechteren Ergebnisse zeitigen.

Beachtet man, dass die Ermittlung des Zugewinns sich auf ein Panel aus Schüler*innen bezieht, die in zwei Jahren den Weg von der achten in die zehnte Klasse schaffen, sollte die Angabe aus der Ermittlung der Messung mit den 16 Items des Blocks 3 an Schüler*innen, die zu beiden Testzeitpunkten anwesend waren als beste Schätzung angegeben werden, also 63,5 Punkte mit einem 95%-Konfidenzintervall von 46,1 bis 80,9. Die Unterschiede zu den anderen Messungen legen nahe, dass die schlechtere inhaltliche Repräsentation keine größeren Auswirkung auf diese Schätzung hat.

Zugewinn von T2 nach T3 Die Untersuchung des Zuwachses zwischen dem neunten und zehnten Schuljahr (Tabelle 5.7) zeigt Vergleichbares. Auch hier sind die Zuwächse auf der Basis nur der Schülerinnen und Schüler, die zu beiden Testzeitpunkten T2 und T3 anwesend waren (N=252), etwas kleiner, als bei der Feststellung unter Einbezug jeweils aller verfügbaren Schüler*innen (T2: N=1200, T3: N=292). Wieder ist zu erkennen, dass der größere Zuwachs auf einen niedrigeren Ausgangswert zu T_2 zurückzuführen ist. Hier wiederholt sich der Effekt, bei dem die Klasse nach der neunten Jahrgangsstufe verlassenden Schülerinnen und Schüler schlechtere Leistungen aufweisen, was vermutlich auch den Grund für deren Verlassen darstellt. Der Zuwachs für die Gruppe der Schülerinnen und Schüler, die 2012 die neunte und ein Jahr später die zehnte Klasse des gleichen Gymnasiums besuchten, beträgt bei großem Konfidenzintervall rund 24 Punkte. Auch wenn die Zuwächse zwischen den Jahrgangsstufen 8 und 9 sowie 9 und 10 nicht notwendig gleich groß sein müssen, zeigen die Analysen von 8 nach 10 einen etwas mehr als doppelt so großen Zuwachs von ca. 60 Punkten. Dass diese Zuwächse insgesamt etwas kleiner sind, als dies bei Stanat et al. (2019a, S. 201) als Erwartung

Abbildung 5.11.: Kompetenzzuwachs auf Basis direkter Verlinkung

formuliert wird, könnte an der spezifischen Population von Berliner Gymnasiast*innen liegen. Sie sind aber von der Größenordnung her und mit Rücksicht auf die Konfidenzintervalle durchaus plausibel.

In der grafischen Darstellung 5.11 werden die Zuwächse abgebildet. Es ist deutlich zu erkennen, dass die Steigungen etwa parallel verlaufen, d.h. die Zuwächse sind über die Zeit sehr ähnlich. Zwischen den zwei Zuwachsmessungen findet sich aber ein deutlicher Versatz der absoluten Werte. Wenn man auch die zwei Werte im Jahr 2012 nicht wirklich vergleichen kann, weil die erste Messung hier keinen Messwert hat und nur durch die Verbindung der zwei 2011 und 2013 gemessenen Kompetenzstände approximiert wird, so müsste man aber eine Kongruenz der Messungen zum Zeitpunkt T_3 (2013) erwarten. Das diese zwei Messungen aber gleich sind, schließen selbst die großen Konfidenzintervalle aus. Die zwei zur Messung des Zuwachses eingesetzten Messinstrumente bilden den Zuwachs also sehr ähnlich und plausibel ab. Die absolute Höhe des Kompetenzstandes wird aber durch die zwei Messungen unterschiedlich angegeben. Die Messung von Zuwächsen mit identischen Instrumenten gelingt gut. Zweifel bleiben lediglich an der Bestimmung der absoluten Höhe der gemessenen Leistung, die offenbar vom verwendeten Instrument abhängt.

Untersuchung des Zuwachses mit indirekt verlinkten Instrumenten. Die Untersuchung der Differenz der Ergebnisse von 2012 und 2011 auf der Basis aller vorliegenden Daten, also ohne Rücksicht auf die verwendeten Instrumente, zeigt einen Zuwachs, der mit nur 14,7 Punkten (CI 4,6 – 24,9) deutlich kleiner ausfällt (Tabelle 5.8) als bei der Berechnung zuvor mit direkt verlinkten Instrumenten. Der Zuwachs von 2012 auf 2013 ist hingegen mit 36,2 Punkten mehr als doppelt so groß. Für den gesamten Zuwachs ergeben sich demnach summarisch gute 50 Punkte.

Die Differenzierung des Zuwachses nach den drei Gruppen zeigt für den ersten Zeitabschnitt

Tabelle 5.8.: Ergebnisse des Kompetenzzugewinns über alle drei Testzeitpunkte

	Instrument	N	Mw	Sd	Se	CI	Diff	CI
T1	2011	1306	521,6	90,9	38,4	[516,6 – 526,5]	14,7[a]	[4,6 – 24,9]
T2	2012	1200	536,3	91,9	40,7	[531,1 – 541,5]		
T1	2011 (G1)	444	533,3	89,3	38,6	[525,0 – 541,6]	13,1	[-3,8 – 30,0]
T2	2012 (G1)	444	546,4	92,2	40,6	[537,8 – 555,0]		
T1	2011 (G2)	228	531,5	83,2	38,2	[520,7 – 542,3]	14,5	[-7,6 – 36,5]
T2	2012 (G2)	228	545,9	86,5	40,4	[534,7 – 557,1]		
T1	2011 (G3)	216	527,1	83,1	38,1	[516,0 – 538,2]	17,6	[-4,6 – 39,7]
T2	2012 (G3)	216	544,6	82,8	40,2	[533,6 – 555,7]		
T2	2012	1200	536,3	91,9	40,7	[531,1 – 541,5]	36,2[b]	[25,1 – 47,2]
T3	alle	1131	572,4	99,8	39,5	[566,6 – 578,3]		
T2	2012 (G1)	486	544,5	92,6	40,6	[536,2 – 552,7]	42,4	[25,6 – 59,2]
T3	2013 (G1)	486	586,9	96,2	40,0	[578,3 – 595,4]		
T2	2012 (G2)	252	546,8	85,1	40,3	[536,3 – 557,4]	23,8	[1,3 – 46,4]
T3	2013 (G2)	252	570,7	97,5	41,2	[558,6 – 582,7]		
T2	2012 (G3)	239	541,3	84,0	40,3	[530,7 – 552,0]	19,3	[-3,7 – 42,4]
T3	2013 (G3)	239	560,6	97,6	36,4	[548,3 – 573,0]		
T1	2011	1306	521,6	90,9	38,4	[516,6 – 526,5]	50,9[c]	[40,1 – 61,6]
T3	alle	1131	572,4	99,8	39,5	[566,6 – 578,3]		
T1	2011 (G1)	495	536,1	88,9	38,7	[528,3 – 543,9]	52,1	[35,6 – 68,5]
T3	2013 (G1)	495	588,2	97,5	40,1	[579,6 – 596,8]		
T1	2011 (G2)	254	531,3	82,1	38,1	[521,2 – 541,4]	40,6	[18,3 – 62,9]
T3	2013 (G2)	254	571,9	99,0	41,3	[559,7 – 584,1]		
T1	2011 (G3)	242	530,0	82,5	38,1	[519,6 – 540,4]	34,1	[10,9 – 57,3]
T3	2013 (G3)	242	564,1	101,5	36,7	[551,3 – 576,9]		

[a] auf der Basis aller gemeinsamen Fälle mit N=1.056.
[b] auf der Basis aller gemeinsamen Fälle mit N=977
[c] auf der Basis aller gemeinsamen Fälle mit N=991

von 2011 nach 2012 nur geringe Unterschiede, so wie es bei einer Trennung in drei möglichst gleiche Gruppen zu erwarten ist. Der Zuwachs variiert hier zwischen 13,1 Punkten für die Gruppe 1 und 17,6 Punkten für die Gruppe 3. Die Schülerinnen und Schüler aller drei Gruppen erhalten dabei zum Zeitpunkt T_1 (2011) und zum Zeitpunkt T_2 (2012) das jeweils aktuelle Testheft. Demgegenüber fallen die Statusmessungen zum letzten Zeitpunkt T_3 (2013) deutlich weiter auseinander. Hier erhalten die drei Gruppen jeweils andere Testhefte.

Die grafische Darstellung der Werte in der Abbildung 5.12 zeigt dies sehr deutlich. Wenn auch die großen Konfidenzintervalle keine klare Aussage zulassen, so kann man erkennen, dass die Messung mit drei verschiedenen Instrumenten zum Zeitpunkt T_3 offensichtlich auch zu drei differenten Ergebnissen und damit Zuwächsen führt.

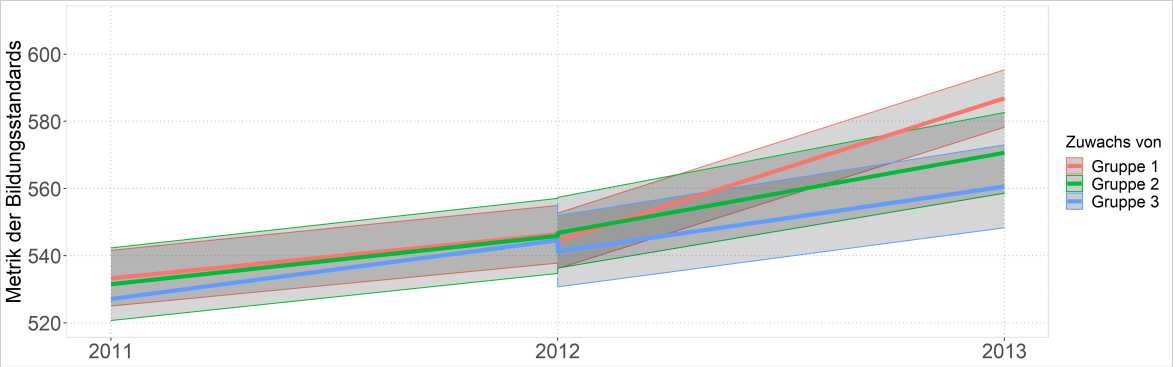


Abbildung 5.12.: Kompetenzzuwachs auf Basis indirekter Verlinkung

Fazit

Die Untersuchungen haben gezeigt, dass längsschnittliche Messungen mit einem identischen Instrument zu plausiblen Aussagen führen. Werden hingegen zwei unterschiedliche Instrumente eingesetzt muss davon ausgegangen werden, dass zwischen den zwei Messungen ein offenbar Instrumenten-bedingter Versatz hinzukommt, über deren Größe keine Aussage getroffen werden kann.

5.5.2. Untersuchung der Stabilität bei mehrfachem Einsatz identischer VERA-Tests

Fragestellung

Mit den folgenden Ausführungen soll die Stabilität bei der Nutzung eines VERA-Instruments im Rahmen einer Panelstudie beurteilt werden. Ein solcher Einsatz wird als stabil angesehen, wenn er die erwarteten Zuwächse zeitigt. Hierbei wird auf einige schon berichtete Ergebnisse zurückgegriffen (Vettorazzi & Harych, 2019).

Während der jährliche Zugewinn für die verschiedenen Domänen auch in der Grundschule zwischen ca. 60 Punkten für Deutsch Lesen (Bremerich-Vos & Böhme, 2009) und Deutsch Zuhören (Behrens et al., 2009) und ca. 80 Punkten für Mathematik (Reiss & Winkelmann, 2009) liegt, fehlt bei Stanat et al. (2017a, S. 153) ein Beleg für die dort angegebenen ca. 100 Punkte auf der Berichtsskala für die Domäne Orthographie[24]. Mit der Panelstudie zur Entwicklung der Orthographiekompetenz sollte diese Lücke geschlossen werden. Erste Ergebnisse dieser Studie sind in einem Projektbericht von Vettorazzi und Harych (2019) vorgelegt worden.

[24]Von den Autoren wird hier ohne Quelle lediglich auf Daten aus der eigenen Normierungsstudie verwiesen.

Entwicklung der Orthographiekompetenz von Klassenstufe 3 zu 4

In diesem Abschnitt wird über eine klassische Panelstudie auf der Basis von VERA-Instrumenten berichtet. Bei VERA-3 wird von Beginn an im Fach Deutsch neben der Domäne *Lesen* eine weitere aus den drei Domänen *Orthographie*, *Zuhören* sowie *Sprache und Sprachgebrauch untersuchen* rotierend angeboten (vergleiche Abbildung A.2). Der Test der Domäne *Orthographie* wurde 2010, 2014 und 2017 im Land Berlin jedes Mal verpflichtend durchgeführt. Zum Zeitpunkt der Durchführung 2010 fehlte als Voraussetzung für die Berichterstattung ein von der KMK bestätigtes Kompetenzstufenmodell für die Domäne Orthographie. Für die letzten zwei Zeitpunkte liegen Ergebnisse als Verteilung der Leistungen der Schülerinnen und Schüler auf Kompetenzstufen vor. Die Ergebnisse des Jahres 2014 sind dem entsprechenden Bericht (Holz et al., 2014) zu entnehmen. Für das Jahr 2017 liegen die Ergebnisse als Antwort auf eine schriftliche Anfrage im Abgeordnetenhaus von Berlin vor (Drucksache, 18/13161, 2018). Der Anteil an Schülerinnen und Schülern, deren Leistungen die Kompetenzstufe I nicht überschreiten und damit den Mindeststandard der KMK nicht erreichen, lag bei 50 bzw. 48 Prozent und damit beide Jahre deutlich über den entsprechenden Werten für die Domäne Deutsch Lesen (26 bzw. 30%). Dies hatte erwartbare öffentliche Reaktionen zur Folge, auch wenn der Bericht (Holz et al., 2014) deutlich ausführt, dass es für diese Differenz fachdidaktische Gründe gibt:

> Außerdem ist zu beachten, dass es sich bei dem Testheft zum Inhaltsbereich Rechtschreiben um ein eher „leichtes“ Testheft handelt. Das heißt, es enthält überproportional viele Aufgaben auf dem Niveau der Kompetenzstufen I und II, da für die Entwicklung der Rechtschreibkompetenz im Mittel ein deutlich stärkerer Kompetenzzuwachs vom Ende der dritten zum Ende der vierten Jahrgangsstufe zu erwarten ist als beispielsweise für Lesen. (ebenda, S.25)

Hier soll der Grund für diese zuerst langsamere Entwicklung der Orthographiekompetenz nicht weiter diskutiert werden. Ansätze dazu finden sich allgemein im Versuch der Ableitung von Stufen der Entwicklung bei Naumann (2008) und mit Verweis darauf konkret für die Vergleichsarbeiten bei Bremerich-Vos und Krelle (2017).

Methode

Für die Untersuchung der Entwicklung der Orthographiekompetenz wurde das Instrument aus der für Berliner Schulen verpflichtenden VERA-Testung vom 4. Mai 2017 (Testzeitpunkt

T_1), im Folgejahr parallel zu VERA-3 als freiwilliger Test für die dann Viertklässler*innen angeboten (Testzeitpunkt T_2). Aus administrativen Gründen war eine Verpflichtung selbst für eine Stichprobe nicht möglich. Eine solche nicht-probabilistische Stichprobe würde die Aussagekraft der Untersuchung aber massiv beeinträchtigt. Deshalb wurde folgendes Vorgehen gewählt:

1. Berechnung der Stichprobengröße.
2. Ziehung einer geschichteten Stichprobe von Schulen, inkl. vonErsatzschulen, aus allen Teilnehmenden des Zeitpunkts T_1.
3. Rekrutierung für T_2, d.h. Bitten der nicht ohnehin freiwillig teilnehmenden Stichprobenschulen um Teilnahme.
4. ggf. Nachziehen von Ersatzschulen.

Im Land Brandenburg war schon die Teilnahme an der Überprüfung der Kompetenzerreichung in der Domäne Orthographie im Rahmen von VERA 2017 freiwillig. Aus diesem Grund wurden als Grundgesamtheit die Berliner Schülerinnen und Schüler gewählt, die zum Zeitpunkt T_1 an öffentlichen Schulen im dritten Jahrgang unterrichtet wurden. Für ein 95%-Konfidenzintervall mit einer Zielgröße von ± 5 Punkten ergibt sich für die Stichprobegröße ein Wert von knapp 400 Schülerinnen und Schülern. Auf der Basis der Vollerhebung der Orthographiekompetenz zu T_1, wurde eine geschichtete Stichprobe aus Lerngruppen an Berliner Grundschulen gezogen. Dabei wurden aus ökonomischen Gründen nur solche Lerngruppen berücksichtigt, die zum Zeitpunkt T_1 mindestens 5 Teilnehmende aufwiesen. Geschichtet wurde zweifach, zuerst nach dem Lerngruppentyp. Ca. ein Drittel der Berliner Lerngruppen wurde 2017 jahrgangsübergreifend unterrichtet (JüL), die anderen jahrgangsbezogen (JbL). Innerhalb des Lerntyps wurden für die Stichprobenziehung aus den Ergebnissen der Vollerhebung zum Zeitpunkt T_1 Leistungsquartile gebildet. Je Lerngruppentyp sollten die Leistungsquartile in der Stichprobe bezogen auf die Anzahl der Schülerinnen und Schüler – nicht auf die Anzahl der Lerngruppen – gleich besetzt sein. Einige der zufällig gezogenen Lerngruppen befanden sich bereits zum Zeitpunkt der Ziehung in der Liste der freiwillig am Messzeitpunkt T2 teilnehmenden Schulen, andere wurden telefonisch und per E-Mail über den Zweck der Studie und die Zufallsauswahl informiert und um ihre Teilnahme ersucht. Die nach nur 4 Absagen[25] und Nachziehungen realisierte Stichprobe wird in Tabelle 5.9 dargestellt. An Hand

[25] Tatsächlich mussten aus der Gruppe der jahrgangsübergreifend unterrichteten Schüler*innen jeweils eine Lerngruppe aus dem dritten und vierten Quartil und aus den jahrgangsbezogenen Lerngruppen je eine aus

Tabelle 5.9.: Stichprobe der Untersuchung zur Kompetenzentwicklung Rechtschreiben

Typ der Lerngruppe		Leistungsquartil				Summe
		1	2	3	4	
jahrgangsübergreifend	Schüler*innen	14	12	12	13	51
	Lerngruppen	2	2	2	2	8
jahrgangsbezogen	Schüler*innen	71	71	64	76	228
	Lerngruppen	4	4	4	4	16
Summe	Schüler*innen					333
	Lerngruppen					24

Tabelle 5.10.: Gegenüberstellung wesentlicher Merkmale der Grundgesamtheit, aller Schüler*innen beider Erhebungen und nur der Stichprobe

Merkmal	Berlin[a]	alle T_1 & T_2[a]	Stichprobe
Anzahl Schüler*innen	25.995	1.245	333
Anteil weiblich	48,6%	47,4%	45,3%
Anteil nicht deutsche Verkehrssprache	32,9%	28,7%	30,3%
Anteil unterrichtet in JüL	15,3%	15,2%	15,3%
Mittelwert	-6,7	-19,8	395,0
Standardabweichung	133,8	128,2	126,2
Mittelwert JüL	-6,7	-38,6	391,6
Mittelwert JbL	-5,7	-16,7	395,7

[a]Mittelwerte sind relativ zum Wert der Stichprobe dargestellt.

der bei den Vergleichsarbeiten standardmäßig erfassten Merkmale zeigt sich eine angemessene Repräsentanz der Grundgesamtheit in der Stichprobe (vergleiche Tabelle 5.10).

Die Panelstudie war in ein Angebot des ISQ eingebunden, Lehrkräften die Wiederholung des Rechtschreibtests aus dem Jahr 2017 bei ihren Viertklässler*innen zu ermöglichen, um so eine standardisierte Einschätzung des Kompetenzzuwachses über das vergangene Schuljahr zu erhalten. Das bedeutet, dass neben den 24 Lerngruppen, die Teil der Stichprobe sind, auch andere teilgenommen haben, inkl. solcher des Landes Brandenburg. Deshalb liegen deutlich mehr Daten vor, als in die Stichprobe zur Feststellung des Zuwachses eingehen (siehe Tabelle 5.10, Spalte *alle* T_1 *&* T_2). Es wäre demnach möglich, die Basis für die Ermittlung des Zuwachses zu erweitern. So könnten alle Lerngruppen, die an beiden Testzeitpunkten teilgenommen haben einbezogen und so gewichtet werden, dass ihr Anteil in der Stichprobenschicht gleich bleibt. Während durch die zufällige Ziehung innerhalb der acht Cluster unbeobachtete Parameter im besten Fall zufällig repräsentiert sein sollten, würde der ggf. nicht zufälligen

dem zweiten und dritten Quartil der gezogenen Ersatzschulen verwendet werden. Die Rekrutierung der Ersatzschulen war erfolgreich.

Selbstauswahl auch bei einer gewichteten Berücksichtigung aller Schülerinnen und Schüler eine Bedeutung zugewiesen. Die Untersuchung des Lernzuwachses mit der kleinen Stichprobe fortzusetzen war deshalb die präferierte Option.

In Brandenburg haben im Jahr 2017 (Zeitpunkt T_1) insgesamt 6.279 Schüler*innen aus 174 Schulen einen Rechtschreibtest geschrieben, davon 5.304 aus 139 öffentlichen Schulen. In der Wiederholungsmessung ein Jahr später (Zeitpunkt T_2) fanden sich nur noch 287 Schüler*innen aus 4 privaten und 5 öffentlichen Schulen. Der Abbildung 5.13 können die nach der Erhebung vorliegenden Daten aus Berliner Schulen entnommen werden. 25.995 Berliner Schüler*innen haben 2017 am hier verpflichtendem Rechtschreibtest teilgenommen[26]. Von den Teilnehmenden waren 1.245 Schülerinnen und Schüler aus 96 Lerngruppen von 41 Schulen an der Wiederholung beteiligt. Für 149 Berliner Schüler*innen liegen 2018 erstmals Daten zum Rechtschreiben vor; 65 von ihnen waren zwar 2017 Teil der Grundgesamtheit, hatten aber an der Testung nicht teilgenommen, 84 sind während des Schuljahres in die vierten Klassen hinzugekommen. Von den 1.245 potentiellen Stichprobenschüler*innen wurden 11 aussortiert, die aus 5 Lerngruppen von 3 Schulen kamen, die weniger als 5 Schüler*innen aufwiesen. So standen letztendlich 1.234 Schüler*innen aus 38 Schulen mit 91 Lerngruppen zur Verfügung. Darunter befanden sich 20 Lerngruppen aus der Hauptstichprobe und 4 nachgezogene mit zusammen 333 Schüler*innen. Aus der Tabelle 5.10 erkennt man sehr gut die korrigierende Wirkung der doppelt geschichteten Stichprobenziehung. Zwar werden die Anteile bezüglich Geschlecht, Verkehrssprache und Unterrichtsform auch von der größeren Stichprobe gut repräsentiert, die Leistungsabweichungen sind allerdings erheblich. Dass die Leistungsmittelwerte der gesamten Stichprobe und in den zwei Teilgruppen der Unterrichtsformen etwa 6 Punkte höher liegen, als in der Grundgesamtheit, kann gerade wegen der kongruenten Auswirkung auf die Teilgruppen vernachlässigt werden. Zudem könnte sich hierin auch der aus der VERAMSA-Untersuchung bekannte Effekt ausdrücken, dass zu zwei aufeinanderfolgenden Zeitpunkten anzutreffende Schüler*innen etwas bessere Leistungen zeigen, als alle Schüler*innen des ersten Zeitpunktes, weil der Drop-out überproportional auf Schüler*innen mit schlechten Leistungen zurückzuführen ist. In der selbst-selektierten Stichprobe sind die Leistungsmittelwerte zum Zeitpunkt T_1 für Lerngruppen mit JüL 30 Punkte und für solche mit JbL noch 10 Punkte niedriger. Diese Abweichungen erklären sich aus den von 25% abweichenden Anteilen der vier Leistungsquartile. Für die Analysen wurden deshalb die 333 Schülerinnen und Schüler der Stichprobenziehung einbezogen.

[26] Der Anteil von 10% nicht-Teilnahmen entspricht der erwarteten Quote.

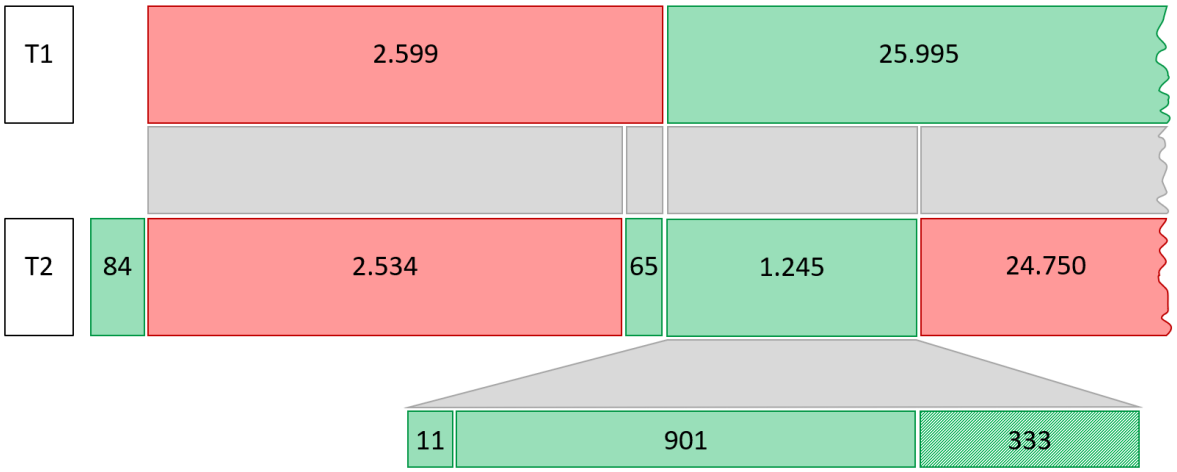


Abbildung 5.13.: Grundgesamtheit zu T_1 und Stichprobenziehung zu T_2 für Berlin

Das Testinstrument wurde vom IQB für die Erfassung des Kompetenzstandes in der Klassenstufe 3 konzipiert. Zwar liegen auch hier, wie bei allen VERA-Tests der Primarstufe, die Bildungsstandards zugrunde, die auf das Ende der Jahrgangsstufe 4 fokussieren, aber mit der Auswahl von Aufgaben wird die mittlere Schwierigkeit des Instruments angemessen gewählt. Gerade wegen des erst *nach* der Testung erwartbaren Kompetenzzugewinns, enthält das Instrument eher leichte Aufgaben. Tatsächlich benötigt man mehr als die Hälfte korrekt gelöster Items, um Leistungen nachzuweisen, welche die in der Kompetenzstufe 1 beschriebenen übersteigen. Es muss befürchtet werden, dass insbesondere für die guten und sehr guten Schüler*innen Deckeneffekte nicht zu vermeiden sind.

Ergebnisse

Die Ergebnisse der zwei Messungen und die sich ergebende Differenz inklusive des Konfidenzintervalls sind in der Tabelle 5.11 wiedergegeben. Diese Ergebnisse sind auch für die zwei Untergruppen *Typ der Lerngruppe* und die *Leistungsquartile* ausgewiesen. Auf Grund der kleinen Stichprobe ist das Konfidenzintervall der Differenz relativ groß, aber das Ergebnis bestätigt mit durchschnittlich 111 Punkten (CI[100;121]) den schon von Stanat et al. (2017b) prognostizierten Zugewinn der Rechtschreibkompetenz von Klasse 3 zu Klasse 4. Zudem stützen die Analysen der Untergruppen die Plausibilität der Ergebnisse. Der Unterschied zwischen jahrgangsübergreifendem und jahrgangsbezogenem Unterricht ist erwartungsgemäß klein. Die Zuwächse der vier Leistungsquartile sind hingegen durchaus unterschiedlich. Insbesondere für die Gruppe der leistungsstarken Schülerinnen und Schüler bestätigt sich der vermutete Deckeneffekt.

Mit der wiederholten Messung der Rechtschreibkompetenz durch die Verwendung eines

Tabelle 5.11.: Ergebnisse der Messungen des Kompetenzzugewinns

	N	Messzeitpunkt T_1			Messzeitpunkt T_2			Differenz $T_2 - T_1$			CI (95%)	
		Mw	Sd	Se	Mw	Sd	Se	Mw	Sd	Se		
	333	395,0	126,2	6,9	505,7	125,6	6,9	110,6	95,4	5,2	100,4	120,9
JüL	51	391,6	111,4	15,6	507,7	121,5	17,0	116,0	100,2	14,0	87,9	144,2
Regel	282	395,7	128,9	7,7	505,3	126,6	7,5	109,7	94,6	5,6	98,6	120,7
Q1	85	317,2	110,3	12,0	451,3	114,3	12,4	134,1	95,0	10,3	113,6	154,6
Q2	83	359,0	109,5	12,0	471,4	123,7	13,6	112,4	85,3	9,4	93,8	131,0
Q3	76	418,2	97,6	11,2	526,5	112,6	12,9	108,3	104,5	12,0	84,5	132,2
Q4	89	483,2	117,3	12,4	571,8	115,2	12,2	88,5	92,6	9,8	69,0	108,0

identischen VERA-Instrument konnte die Stabilität des Instruments trotz der kleinen Stichprobe belegt werden. Insbesondere plausibilisieren auch die Ergebnisse der Teilgruppen das Ergebnis. Für die Plausibilität der Ergebnisse gleichermaßen wichtig ist vermutlich die konservative Stichprobenziehung. Eine umfassende Analyse der Untersuchung zur Entwicklung der Rechtschreibkompetenz ist aktuell in Arbeit.

5.6. Diskussion

Dieses Kapitel untersuchte die Stabilität der Messungen von Kompetenzständen im Rahmen der Vergleichsarbeiten, insbesondere im Lichte mancherorts vorgetragener längsschnittlicher Interpretationen. Letztere sind allerdings auch durch die Implementation der Vergleichsarbeiten intendiert worden. Schon die Untersuchung einiger psychometrischer Aspekte im vorhergehenden Kapitel 4, indizierte einige Schwierigkeiten. Hinzu kommen nun weitere Verwerfungen, die auch in der Auseinandersetzung mit praktischen Messungen im Feld ihren Niederschlag finden.

Veränderungen bei VERA innerhalb eines Jahres, so konnte hier gezeigt werden, sind in der Regel größer oder sogar deutlich größer als Veränderungen beim Bildungstrend über 5 bzw. 6 Jahre. Dies kann kein Abbild wahrer Kompetenzstände sein. Man muss bei den erklärbaren Darstellungen von Einzelaspekten sowohl bei der Auswertung des Bildungstrends wie auch bei den hier ausgewählten Untersuchungen mit Instrumenten der Vergleichsarbeiten vermuten, dass offensichtlich andere Einflüsse diese Veränderungen produzieren. Wie im Abschnitt 5.5 gezeigt werden konnte, messen die Instrumente für sich bzw. miteinander direkt verlinkte Instrumente den Kompetenzstand sehr verlässlich. In einer solchen Konstellation erbringt die Messung eines Zuwachses hochgradig plausible Ergebnisse und deshalb sind auch die Ergebnis-

se des Bildungstrends keinesfalls in Zweifel zu ziehen. Die Messung von Kompetenzständen mit verschiedenen, nur indirekt (über die VERA-Pilotierung) verlinkten Instrumenten der Vergleichsarbeiten zeigen allerdings unplausible Verschiebungen (siehe Abschnitt 5.5.1). Diese Messungen stellen vermutlich ein klassisches Beispiel einer zwar reliablen, aber eben nicht validen Messung dar. Die Ergebnisse stützen aus der Sicht des Verfassers die Vermutung, dass das Linking der VERA-Instrumente im Rahmen der Pilotierung zumindest teilweise für die dargelegten Problemlagen verantwortlich ist. Die mit dieser Bestandsaufnahme aufgeworfenen Probleme, rechtfertigten eine tiefer gehende Untersuchung. Bei einer Bestätigung der hier vorgetragenen Befunde, sind Prozessänderungen unabdingbar. Denn das Problem ist nicht ein Versatz, der in einem Jahr 30 Punkte mehr und im nächsten 10 weniger anzeigt, sondern dass der Bezug zu den Kompetenzstufenmodellen damit beliebig wird. Dieser kriterielle Bezugspunkt ist aber die Essenz der Rückmeldungen, sowohl auf individueller Ebene, für die Lehrkraft und für die Schule, wie auch aggregiert für das Land.

Natürlich sollte den hier beschriebenen Instabilitäten mit einer Untersuchung begegnet werden, deren Ziel es ist, kompetenzstufenbasierte Rückmeldungen im Kontext von VERA eine angemessen valide Grundlage zu geben. Da aus politischen Gründen ein Aussetzen von VERA ausgeschlossen ist, deuteten Harych und Emmrich (2014) einen pragmatischen Vorschlag an, der grobe Unstimmigkeiten verhindern kann. Dabei wird die Skalierung nach der Durchführung der Vergleichsarbeiten wie im technischen Manual beschrieben regelkonform durchgeführt. Allerdings wird das Ergebnis anschließend derart landesspezifisch normiert, dass der Mittelwert in einem plausiblen Verhältnis zur letzten Messung beim Bildungstrend steht. So ergab sich beim letzten Bildungstrend für Berliner Neuntklässler*innen im Fach Mathematik ein Mittelwert von 479 Punkten. Bei einem jährlichen Zuwachs von ca. 50 Punkten (Stanat et al., 2019a), erwartet man im Rahmen der Vergleichsarbeiten einen Mittelwert von ca. 429 Punkten. Der mit den Vergleichsarbeiten 2020 zuletzt festgestellte Wert liegt allerdings bei 492 Punkten, also 63 Punkte höher. Deshalb werden sämtliche Werte der Kompetenzmessung für das Fach Mathematik bei VERA um diese Differenz korrigiert. Eine echte Veränderung von Kompetenzständen zwischen zwei Jahres-Kohorten des Landes ist damit nicht mehr feststellbar, weil zu Null normiert. Allerdings lassen die obigen Ausführungen kaum Zweifel daran, dass eine Feststellung genau dieser Veränderung aktuell nicht valide ist. Mit dem skizzierten auch als Mean-Mean-Equating[27] bekannten Verfahren werden die Ergebnisse der Schulen über die Jahre vergleichbar. Auch vergangene Ergebnisse können mit

[27] Dabei werden alle Wert einer Erhebung so transformiert, dass die verschiedenen Mittelwerte zweier Messungen angeglichen werden.

dieser Methode vergleichbar gemacht werden. Die Bezüge zu den Kompetenzstufen, deren inhaltliche Beschreibung Teil der Rückmeldungen ist, werden damit – im vorliegenden Fall für das Jahr 2020 dramatisch – verändert. Es muss aber angenommen werden, dass sie in den Jahren zuvor in ähnlichen Größenordnungen Verzerrungen unterworfen waren, die sich kaum erklären ließen. Schlimmer noch, führten sie in Schulen und bei Lehrkräften vielleicht zum Versuch, diese Veränderungen zu erklären und mit Maßnahmen zu bearbeiten, obwohl sie keine oder nur teilweise reale Anteile hatten. Tatsächlich empfahl auch das IQB ein solches Mean-Mean-Equating (Weirich, 2016), als die Steigerung der Anteile der Kompetenzstufe V bei der Feststellung der Lesekompetenz in der Primarstufe 2016 (vergleiche Abschnitt 5.4.2) von mehreren Ländern als unplausibel reklamiert wurde. Die darin unterstellte Annahme, dass sich die tatsächlichen Kompetenzstände innerhalb eines Jahres nur wenig ändern, belegen die Zuwächse im Bildungstrend hinreichend, wie auch hier gezeigt werden konnte.

Trotzdem bleibt festzustellen, dass keine der hier referierten Untersuchungen das Ziel hatte, die Stabilität der Messung bei Vergleichsarbeiten zu prüfen. Eine solche Untersuchung fehlt bislang und die Ausführungen haben deren Notwendigkeit hinreichend belegt. Dieses Kapitel ist in diesem Sinne eine Bestandsaufnahme und Vorbereitung einer solchen Untersuchung.

6. Vor der Rezeption

6.1. Theoretischer Hintergrund

Die Überschrift dieses Kapitels bezieht sich auf das vermutlich im Rahmen der Vergleichsarbeiten am häufigsten zitierte Modell für die Ergebnisnutzung von Helmke (2004). Darin werden vier Prozessschritte der Nutzung *Rezeption*, *Reflexion*, *Aktion* und *Evaluation* sowie die auf sie einwirkenden individuellen und externen Rahmenbedingungen unterschieden (reduzierte Abbildung 6.1). Vermutlich wegen der als eher randständig wahrgenommenen Bedeutung wird der Abruf von Rückmeldungen als Teil der *technischen Übermittlung* der Daten neben der Auswahl[1] und dem Verständnis der Rezeption zugeordnet, vielleicht sogar chronologisch *vor*geordnet. Wenngleich diese Übermittlung eine basale Grundlage für die eigentliche Rezeption darstellt, ist sie davon zu differenzieren und gerade nach trivial zu operationalisieren. Nachdem die Vergleichsarbeiten mehr als 15 Jahre verpflichtender Bestandteil der Primarstufe und mehr als 10 Jahre der Sekundarstufe I sind, kann sicher davon ausgegangen werden, dass der technische Prozess des Abrufs sowohl auf Grund gestiegener EDV-Kompetenz der Lehrkräfte wie auch der weiter entwickelten Portale kein nennenswertes Problem darstellt. Ob eine Rückmeldung abgerufen wird, kann damit der Konstellation individueller und externer Bedingungen zugeschrieben werden. Mit der *Motivation* ist eine individuelle Bedingung im Modell explizit benannt.

Bei Groß Ophoff (2013, S. 56) findet sich eine Auseinandersetzung mit verschiedenen Theorien der Motivation in Verbindung mit der Nutzung von Ergebnisrückmeldungen aus Vergleichsarbeiten. Sie übersetzt zum Beispiel die entlang einer kontinuierlichen Skala von Selbstbestimmung angeordneten Motivationstypen des Modells der Selbstbestimmungstheorie von Deci und Ryan (2000) für die Ergebnisnutzung durch Lehrkräfte. Darüber hinaus macht sie deutlich, wie sich die funktionelle Bedeutung auf die Selbstregulation auswirkt und dass „die zur Verfügung gestellten Rückmeldungen in den Vergleichsarbeiten von den beteiligten Lehrkräften idealerweise als informational wahrgenommen werden sollten, um wie

[1] In späteren Veröffentlichungen findet sich hier statt *Auswahl* der Begriff *Aktualität*.

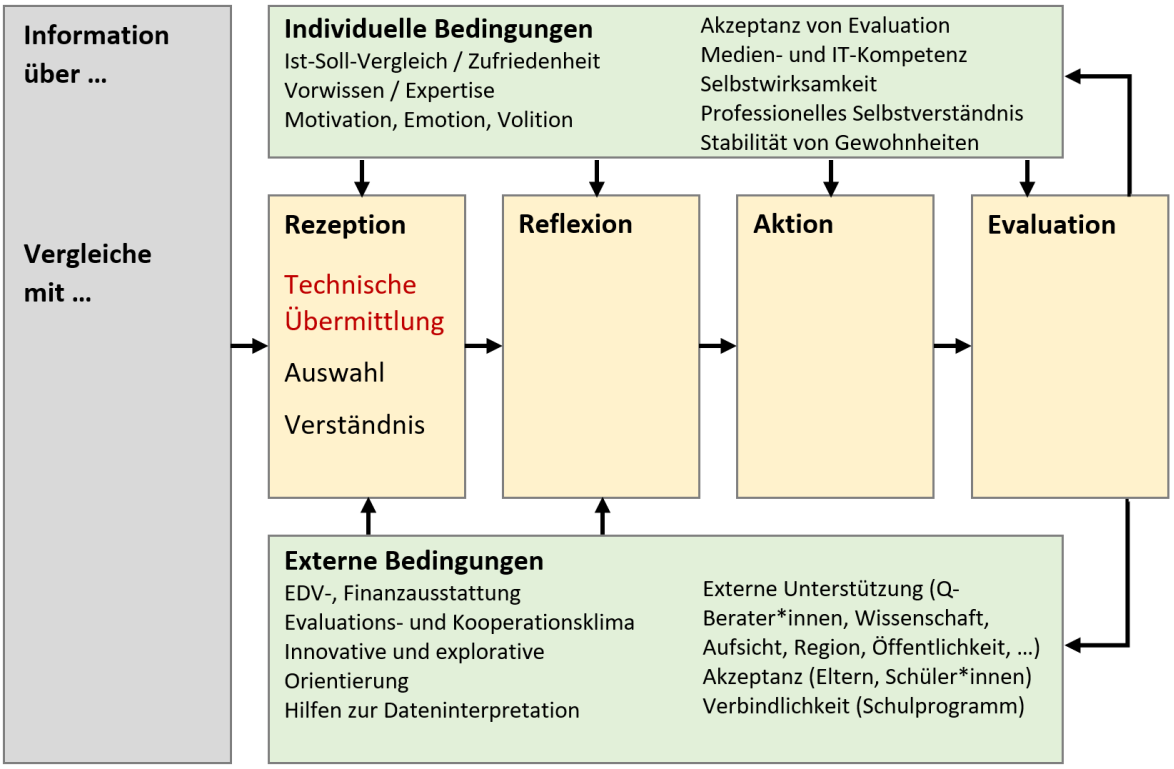


Abbildung 6.1.: Rahmenmodell zur Ergebnisnutzung der Vergleichsarbeiten nach Helmke (2004, S. 11), mit verkürzter inhaltlicher Beschreibung

intendiert zur Qualitätsentwicklung genutzt zu werden.“

Tatsächlich gibt es aber Anlässe, die den Abruf einer Rückmeldung unabhängig von diesem intendiertem Interesse herausfordern, und die sich durch die Erfassung des Abrufs allein nicht unterscheiden lassen. Die Herausgabe einer individuellen Rückmeldung an jede Schülerin und jeden Schüler bzw. an deren Eltern ist für die Länder Berlin und Brandenburg als obligatorisch festgelegt und die Schulen erhalten für die Erziehungsberechtigten jedes Kindes einen Elternflyer, der diese Obligation ausweist. Zudem verlangt die Schulinspektion zumindest in Berlin im Vorgang ihres Schulbesuchs die Übergabe verschiedener Unterlagen durch die Schule, darunter auch VERA-Rückmeldungen der zurückliegenden drei Jahre. Überdies sind partiell Anlässe denkbar, bei denen die Rückmeldungen der Vergleichsarbeiten eine Rolle spielen könnten, wie zum Beispiel im Rahmen der Bilanzgespräche zwischen Schulleitung und Schulaufsicht. Immerhin gibt es in der Grundschule nur wenige standardisierte Parameter zur Beurteilung der von den Schülerinnen und Schülern erreichten Fachleistungen. Ein Abruf kann ggf. aber auch nur dazu erfolgen, die Rückmeldung zu speichern, auszudrucken oder abzulegen, ohne das sich eine intendierte Aktion anschließt. Vielleicht wird aber mit der dichotomen Variable des Abrufs einer Rückmeldung weniger vielschichtig kausal begründbare Motivation operationalisiert als eher Amotivation. Aber auch hier wäre eine Identifikation

mit dem parallel dazu definierten Typ im Modell von Deci und Ryan (2000) zu weitgehend. Wenn beispielsweise Kritiker*innen der Vergleichsarbeiten die VERA-Ergebnisse als für ihren Unterricht irrelevant ablehnen und sie deshalb nicht herunterladen, handeln diese hochgradig selbstbestimmt.

Die Nutzungsforschung hat der Tatsache, dass Rückmeldungen ggf. nicht genutzt werden bisher wenig Aufmerksamkeit gewidmet. Dieses Kapitel bereitet die Daten der Rückmeldeabrufe auf und untersucht sie anhand ausgewählter Forschungsfragen. Tatsächlich muss wegen der weitgehend verpflichtenden Durchführung von VERA erwartet werden, dass der größte Teil der Rückmeldungen abgerufen wird. Trotzdem kann dann der Frage nachgegangen werden, zu welchen Zeitpunkten bzw. wie schnell Rückmeldungen abgerufen werden.

Der erste Teil dieses Kapitels präsentiert einige wesentliche Forschungsarbeiten zur Ergebnisnutzung der Vergleichsarbeiten der letzten zwei Jahrzehnte. Hier stehen in erster Linie Fragen des Verständnisses der Testkonstruktion inkl. der Durchführung, der zugrundeliegenden fachdidaktischen Konstrukte, der statistischen Verfahren der Analyse, der Darstellung der Ergebnisse und deren fachdidaktische Interpretation und die Nutzung des gewonnenen Wissens für die Gestaltung von Unterricht sowie von Prozessen der Unterrichtsplanung und -durchführung in der Schule im Mittelpunkt. Diese Forschungen sind oft quantitativ und beziehen sich vielfach auf über Fragebögen erhobene Selbsteinschätzungen. Die Vergleichsarbeiten erfahren durch die bundesweite Verpflichtung einerseits und als im Rahmen von Reformprozessen top-down implementiertes Werkzeug andererseits besondere Aufmerksamkeit. Es kann erwartet werden, dass insbesondere anonym erhobene Selbsteinschätzungen davon nicht unbeeinflusst bleiben und so stellt sich immer auch die Frage nach der Verlässlichkeit solcher Aussagen. In letzter Zeit finden sich in Forschungsarbeiten häufiger qualitative Ansätze, welche die Prozesse tiefgreifender offenzulegen versuchen. Hier geht es allerdings selten um die Akzeptanz des Verfahrens, sondern um konkrete Nutzungsprozesse, oft im Zusammenhang mit konkreten Rückmeldungen.

Die anschließende Übersicht zum Forschungsstand untersucht die entwickelten Modelle und Forschungsansätze in Hinsicht darauf, ob sie auf den technischen Prozess des Abrufens von Rückmeldungen rekurrieren bzw. diesen wenigstens implizit adressieren. Es konnten keine Forschungsvorhaben recherchiert werden, die sich mit dem Abruf der Rückmeldungen beschäftigt haben. Ungeachtet dessen ist der leicht operationalisierbare Abruf einer Rückmeldung grundlegend für alle Folgeprozesse und damit ein „harter“ Indikator für die Nutzung. Im Abschnitt 6.3 sind wegen fehlender theoretischer Grundlagen keine Hypothesen, sondern

lediglich auf Erfahrungen im Praktischen zurückgehende Forschungsfragen formuliert.

6.2. Forschungsstand

Insbesondere in den ersten Jahren nach der Implementation der Vergleichsarbeiten wurden viele Untersuchungen zur Rezeption von Rückmeldungen durchgeführt, die in verschiedenen Quellen gut zusammengefasst wurden, so schon bei Koch et al. (2006), später bei Dedering (2011) oder Groß Ophoff (2013) und in einem kritischen Rückblick auf 10 Jahre VERA durch Zimmer-Müller et al. (2014) oder jüngst von Pukrop (2019, S.32-56). Die Mehrzahl der rezipierten Untersuchungen bezieht sich dabei auf verschiedene Modelle, welche die Rezeption von Rückmeldungen, oft aber den gesamten Prozess eines Evaluationskreislaufs mit der Durchführung der Vergleichsarbeiten als einen Teilprozess beschreiben. Eine umfassende Übersicht über verschiedene Modelle findet sich wieder bei Pukrop (2019, S.62-84). Altrichter et al. (2016) setzt sich kritisch mit solchen Modellen und deren fehlender Explikation von Prozessen auseinander:

> „Der kritische Punkt in derlei Modellen ist aber, dass die ‚Pfeile' [...] mehr den Zusammenhang zwischen einzelnen Akteuren, Elementen und Ebenen beschwören, als dessen genaue Qualität explizieren, dass sie oft suggerieren, ein solcher Zusammenhang müsse sich notwendig ergeben, wo er in der Realität durchaus mühsam herzustellen und prekär in seiner Aufrechterhaltung ist." (ebenda S.239)

Auch das Helmke-Modell beschreibt idealtypische Zusammenhänge und Prozesse. Nicht jede reale Situation wird mit diesem Modell vollständig abgebildet und nicht jedes Modelldetail hat in einem konkreten Fall eine reale Widerspieglung. Die Bildunterschrift „Der Zyklus Rezeption – Reflexion – Aktion - Evaluation"[2] lässt die Interpretation wiederkehrender Teilprozesse in einem Evaluationskreislauf zu, bedingt diese aber nicht notwendig. Der detailreichen Untersetzung der Teilprozesse ist es offenbar geschuldet, dass die *Technische Übermittlung* der zu rezipierenden Rückmeldung als solche überhaupt Erwähnung findet. In vielen anderen Modellen ist dies vorausgesetzt und bleibt wohl deshalb unerwähnt. Ursächlich für die Erwähnung der technischen Übermittlung bei Helmke (2004) ist vermutlich auch, dass diese Prozesse für die damals ersten VERA-Durchgänge an Grundschulen einiges technisches Können auf Seiten der Lehrkräfte flächendeckend voraussetzen mussten.

Einige der Veröffentlichungen zur Rezeptionsforschung sollen nun dahingehend untersucht

[2] So zu finden in der ersten publizierten Version bei Helmke (2004)

werden, ob sie den Abruf von Rückmeldungen betrachten, oder sogar quantifizieren, bzw. ob sich aus den Auswertungen Schlussfolgerungen dazu ziehen lassen. Die Veröffentlichungen werden dazu in drei Gruppen aufgeteilt: Der erste Abschnitt bezieht sich auf solche Veröffentlichungen, die keine Rückschlüsse auf Abrufquoten zulassen. Sie sind hier dennoch aufgeführt, weil erkennbar wird, dass die Abrufquoten die Ergebnisse vermutlich beeinflusst haben. Eine zweite Gruppe von Veröffentlichungen lässt mittelbare Rückschlüsse auf die Abrufquoten von Rückmeldungen zu. Es verwundert, dass für die dritte Gruppe mit quantitativen Aussagen zu den Abrufen von VERA-Rückmeldungen lediglich Berichte aus einer Quelle in Thüringen zu finden waren. Da Rückmeldungen heute mehrheitlich, wenn nicht gar vollständig digital verfügbar gemacht werden, kann eine Abruferfassung leicht von Webapplikationen realisiert werden.

6.2.1. Veröffentlichungen, die keine Rückschlüsse auf Abrufquoten zulassen

Aus vielen Studienberichten lassen sich keine Daten oder Hinweise zum Abruf oder zur Inaugenscheinnahme von Rückmeldungen entnehmen. So wird beispielsweise bei Kuper und Diemer (2012) schlicht davon ausgegangen, dass dies passiert (ist) bzw. wurden keine Personen zu Interviews rekrutiert, die sich nicht mit den Rückmeldungen auseinandergesetzt haben.

Hilfreich können hier also lediglich solche, zuvorderst quantitative Untersuchungen sein, die dem Anspruch von Repräsentativität zu entsprechen versuchen. Immerhin können zum Beispiel aber Interviews Hinweise darauf geben, warum ein Zugriff auf Rückmeldungen nicht stattfindet. So interviewte Jäger (2012) Lehrkräfte aus Schulen verschiedener Schulformen des Landes Baden-Württemberg zum Umgang mit der Rückmeldung aus Vergleichsarbeiten, wobei sich der Leitfaden dicht am Helmke-Modell orientierte. Unter der Hauptkategorie II des Bereichs Rezeption wurde die Verständlichkeit der Rückmeldungen in drei Subkategorien kodiert, wobei neben „Rückmeldungen waren verständlich“ und „Rückmeldungen waren unverständlich/ unpraktisch“ auch „Rückmeldungen wurden nicht angeschaut“ dezidiert erfasst wurde. Das meint mindestens jene Lehrkräfte, welche die Rückmeldungen in der Applikation nicht angeklickt hatten, zusätzlich aber auch jene, die das getan, sich die Rückmeldungen dann aber trotzdem nicht angesehen haben. Von den 59 interviewten Lehrkräften gaben 4 der 21 Gymnasiallehrkräfte und 9 der 38 Lehrkräfte aus Haupt- und Realschulen und damit in insgesamt 22% an, sich die Rückmeldungen nicht angeschaut zu haben (ebenda, S.158). Antworten von Lehrkräften dieser Kategorie wurden teilweise als „Eingeständnisse“ bewertet,

es wurden aber auch verschiedene Gründe angeführt, wie fehlende Zeit, Skepsis gegenüber der Aussagekraft oder schlicht Desinteresse. Wer meint, dass bestimmte Quellen qualitativ bessere Informationen bereitstellen, drückt damit spezifisches Interesse für etwas und damit gleichzeitig ein geringeres, oder eben Desinteresse für die Alternative aus (ebenda, S.160). Die Stichprobe ist für eine Interviewstudie umfangreich. Die dargestellte Ziehung ist nachvollziehbar, führt allerdings, so muss angenommen werden, in der Konsequenz zu einer Überschätzung der Abrufe von Rückmeldungen.

Die Autoren von Maier et al. (2012) schlussfolgern am Ende ihrer einführenden Diskussion: „Ob und wie effektiv Vergleichsarbeitsrückmeldungen tatsächlich in Lehrerkollegien rezipiert und für weiterführende Entscheidungen genutzt werden, hängt sowohl vom Testsystem selbst, den Bedingungen der Implementation als auch von den schulischen Unterstützungssystemen ab“ (ebenda S.205), geben damit aber wenig Hinweise für die Häufigkeit von Rückmeldeabrufen. Sie machen allerdings deutlich, dass bei der schulinternen Datennutzung zwischen verschiedenen Ebenen wie der einzelnen Lehrkraft, der Gruppe der Fachkolleg*innen (zum Beispiel als Fachkonferenz) oder der Schulleitung zu differenzieren sei (ebenda S.204).

6.2.2. Veröffentlichungen, die begrenzte Rückschlüsse auf Abrufquoten zulassen

Konkret wurden von Bach et al. (2014) Brandenburger a) Schulleitungen (tatsächlich sind es 23 Leitungen von Gymnasien und 36 von weiterführenden Schulen, sowie überwiegend 104 Grundschulen) und b) Lehrkräfte (70 Lehrkräfte von Gymnasien und 105 von weiterführenden Schulen) zu konkreten Aspekten der Nutzung von Rückmeldung befragt. Auf der Basis einiger erhobener Variablen wurde Repräsentativität angenommen. 48% der Schulleitungen gaben an, auf der Basis der Rückmeldungen Entwicklungsprozesse im Bereich der Personalentwicklung angestoßen zu haben (ebenda, S.73). Zur Erklärung dafür erwiesen sich weder die Qualifikation, noch das Vorhandensein einer erweiterten Schulleitung oder das Alter als signifikant, einzig die wahrgenommene Nützlichkeit der VERA-Rückmeldungen. Der Mittelwert der 6stufigen Skala der wahrgenommenen Nützlichkeit liegt mit 3,9 (SD=1,32) 0,4 Punkte über dem nominellen Skalenmittelwert. Es wurden an keiner Stelle schulformspezifische Zahlen berichtet, so dass der Mittelwert wie die Effekte in der logistischen Regression ggf. von dem mit 64% großen Anteil Grundschulleitungen überzeichnet worden sein könnte. Die Schulleitungen bestätigten hingegen mit 82% eine häufigere Nutzung der Daten für Unterrichtsentwicklung, 73% konkretisierten dies mit der Ableitung von Fortbildungsmaßnahmen. Aus Sicht der Schulleitung werden die VERA-Rückmeldungen weniger auf der Ebene der

Schule, denn der Klasse bzw. im Rahmen der Fachkonferenz genutzt.

Ein Drittel geben allerdings an, dass die Daten in den Fachkonferenzen überhaupt nicht genutzt werden (ebenda S.74), umgekehrt geben also zwei Drittel an, dass die VERA-Rückmeldungen in irgendeiner Form dort genutzt werden. Lehrkräfte bewerten die Nützlichkeit hingegen mit 2,86 (SD=1,35) deutlich schlechter als Schulleitungen. Wollte man auf der Basis dieser Zahlen eine hilfreiche Abschätzung dafür abgeben, wie viele Schulleitungen den Link für die VERA-Rückmeldungen mindestens angeklickt haben, muss aber berücksichtigt werden, dass für etwa ein Viertel der Schulleitungen bestimmte „relevante Merkmale“ (ebenda S. 69) nicht vorlagen, die deshalb keinen Eingang in die Untersuchung gefunden haben. Dies könnten zumindest teilweise genau solche Schulleitungen sein, welche die Rückmeldung erst gar nicht in Augenschein genommen haben. Unabhängig davon beteiligten sich ohnehin nur 34% der Schulleitungen.

Richter et al. (2014) konnten in ihrer Studie nachweisen, dass Lehrkräfte, die davon überzeugt sind, dass die Vergleichsarbeiten als Mittel zur Unterrichtsentwicklung dienen können, die Ergebnisse von Vergleichsarbeiten zur Unterrichtsentwicklung einsetzen, wobei aber unklar bleibt, wie viele Lehrkräfte dies denn tatsächlich betrifft. Unabhängig davon wurden auch 12 Fragen dazu gestellt, zu welchen Veränderungen im Unterricht VERA führt, die zu vier Skalen verdichtet wurden. Drei beziehen sich auf konkrete Veränderungen, während die vierte Skala „keine Veränderung“ zu messen vorgab. Der Mittelwert dieser aus zwei Fragen gebildeten Skala weist mit 2,71 als einzige der vier erhobenen Skalen in den zustimmenden Bereich, wobei hier der Aussage, dass „keine Veränderung“ stattgefunden hat, zugestimmt wurde. Dabei verweist das im Bericht exemplarisch aufgeführte Item in seiner Aussage aber nicht einfach auf „keine Veränderung“ oder wie im Ergebnisteil formuliert auf eine „Konstanz in der Unterrichtsentwicklung“ (ebenda S.14), sondern deutlich schärfer mit „Ich halte es für falsch, wegen Leistungsvergleichen Veränderungen in meinem Unterricht vorzunehmen.“ (ebenda S.23) auf eine klare Ablehnung. Der Grad der hier formulierten Ablehnung ist offensichtlich erheblich. Leider fehlen hier konkrete Zahlen, die ein so ausgedrücktes Desinteresse quantifizieren.

Es kann nicht ausgeschlossen werden bzw. scheint es sogar plausibel, dass die freiwilligen Stichproben genau solche Schulen ausgeschlossen haben, welche die Rückmeldung nicht heruntergeladen haben. Allerdings sind auch schwerlich Untersuchungssettings mit freiwilliger Rekrutierung vorstellbar, die hierfür eine gute Operationalisierung finden. Selbst bei quasi verpflichtenden Befragungen, scheinen solche Aussagen wenig verlässlich, um eine Abschät-

zung des fehlenden Abrufs von Rückmeldungen zu erlauben. Schulleitungen wissen, dass die Nutzung von VERA-Ergebnissen erwartet wird und würden es deshalb vermutlich vorziehen a) nicht an einer solchen Studie teilzunehmen, b) vor der Teilnahme die Rückmeldung doch kurz zu inspizieren oder c) die Fragen wenigstens im Sinne der sozialen, oder präziser der administrativen Erwünschtheit bzw. der auftragsgemäßen Anforderungen zu beantworten. Demnach ist das Ausmaß von Ablehnung bei Richter et al. (2014) vermutlich eher eine Unterschätzung.

Schon 2004 schränkten Kohler und Schrader (2004) die Belastbarkeit der Aussagen von Befragungen ein und machten zwei Quellen dafür verantwortlich. Zum einen bemängeln sie die Rücklaufquote und vermuten, dass die Bereitschaft zur Teilnahme mit der allgemeinen Einschätzung der Evaluationsstudie konfundiert sein könnte. Zum anderen äußern sie begründete Zweifel an der Verlässlichkeit der Selbstauskünfte, welche die Befragungen zumeist abverlangen. Diese Umstände schränken die Aussagekraft von Untersuchungen ggf. ein.

6.2.3. Veröffentlichungen mit quantitativen Aussagen zu Abrufquoten

Auch deshalb kommt der einzigen Quelle, die konkrete Abrufzahlen berichtet, eine besondere Bedeutung zu. Nachtigall vermutete schon 2005 (Nachtigall, 2005), dass sich die Beteiligung an Vergleichsarbeiten und damit auch die Nutzung der Ergebnisse in drei Phasen entwickeln wird. In der Einführungsphase würde das neue Instrument VERA von allen Seiten inkl. der Administration einer skeptischen Prüfung unterzogen, Nachtigall berichtete von nicht seltenem Misstrauen und von Ablehnung auf Seiten der Lehrkräfte. Nach erfolgreichem ersten Durchlauf folgte im zweiten Jahr eine Phase der Akzeptanz, über die ebenda berichtet wird. Die Zustimmungswerte, die in der obligatorischen Befragung gemessen wurden, erhöhten sich deutlich. Anfängliche Ängste konnten offenbar zerstreut werden und die durchführende Universität, so vermutet Nachtigall, wurde als neutraler Partner erlebt. Vielfältige Schulungen zur Nutzung der Ergebnisse haben diese Entwicklung vermutlich unterstützt. Trotzdem gibt Nachtigall selbst für diese positive Situation zu Protokoll: „Die Anzahl der Schulen, die ihre Ergebnisse zwei Monate nach Lieferdatum noch nicht abgerufen hatten, war mit 32% enttäuschend und inakzeptabel hoch.“ (ebenda S.90). Da jede Schule sowohl mehrere Typen von Rückmeldungen jeweils zu jedem getesteten Fach erhält und weil der Bericht für weiterführende Schulen die Rückmeldungen zu VERA-6 und 8 zusammen behandelt, ist unklar, ob hier 32% der Schulen *gar keine* oder *nicht alle* Rückmeldung abgerufen haben oder ob 32% der Rückmeldungen noch nicht abgerufen worden sind und auch nicht, ob dies für VERA-6

wie 8 gleichermaßen zutrifft. Für das kommende, aus der Perspektive des Berichts dritten VERA-Jahres sieht Nachtigall die Voraussetzungen für eine dritte Phase gegeben, die von einer verstärkten Nutzung der Tests gekennzeichnet sein soll.

Die zeitliche Betrachtung der Rezeption in einer Rückschau (Nachtigall & Hellrung, 2013) zeigt, dass die Akzeptanz-Phase 2005 nicht nur eine Steigerung zum Vorjahr, sondern tatsächlich den Höhepunkt der wahrgenommenen Nützlichkeit auszeichnet und dass diese danach stetig zu sinken begann und inzwischen auf oder gar unter dem Status des ersten Jahres liegt. Die Annahme, dass sich die Abrufzahlen parallel dazu entwickelt haben, lässt sich an Hand der in den jährlichen Landesberichten ausgeführten Downloadquoten untersuchen. In den Landesberichten für das Land Thüringen (Nachtigall, 2008 sowie Nachtigall, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020) findet sich die einzige Quelle, die dezidiert Auskunft über Downloadquoten für Rückmeldungen im Rahmen von Lernstandserhebungen gibt. Auch hier werden sämtliche Vergleichsarbeiten (Jahrgänge 3, 6 und 8) zusammen präsentiert. Die Differenzierung in Schulformen führt dabei zwar zu einer separaten Darstellung der Quoten für VERA-3, für die weiterführenden Schulen sind aber die Rückmeldungen zu VERA-8 und VERA-6 nicht zu trennen. Insbesondere wegen der teilweise freiwilligen Teilnahme an den Vergleichsarbeiten in der Klassenstufe 6, sollte das Abrufverhalten für VERA-8 allein tendenziell eher etwas unter den hier berichteten Werten liegen. Im Anhang findet sich eine tabellarische Zusammenfassung der Abrufzahlen aus den 12 Jahresberichten (A.7.1). Fasst man die Ergebnisse der vorliegenden Landesberichte von 2008 und 2010 bis 2018 zusammen, ergibt sich die Abbildung 6.2 von mehr oder weniger stetig, ganz klar aber von der Tendenz her sinkenden Abrufquoten. Auf Grund der Berichtsform akkumulieren die abgebildeten Werte die Rückmeldeabrufe für VERA-6 und 8.

Für die fachspezifischen Rückmeldungen in Abbildung 6.3 ließen sich die Rückmeldequoten ca. 18 Wochen nach Freischaltung aus einer zweiten Graphik der Landesberichte entnehmen, hier auch spezifisch für die einzelnen Fächer. Erwartungsgemäß sind die Abrufzahlen direkt nach der Freischaltung hoch und sinken dann über die Zeit. Nach ca. 18 bis 20 Wochen werden offenbar nur noch vereinzelt Rückmeldungen abgerufen. Auch hier ist deutlich erkennbar, dass die Abrufquoten über die Jahre fortschreitend sinken. Wie den Berichten und der Graphik 6.3 zu entnehmen ist, liegen die Rückmeldequoten für das Fach Mathematik leicht über denen für Deutsch und Englisch.

Zwischen der abnehmenden wahrgenommenen Nützlichkeit (Nachtigall & Hellrung, 2013) und den gleichermaßen abnehmenden Abrufquoten (Abbildungen 6.2 und 6.3) könnte ein

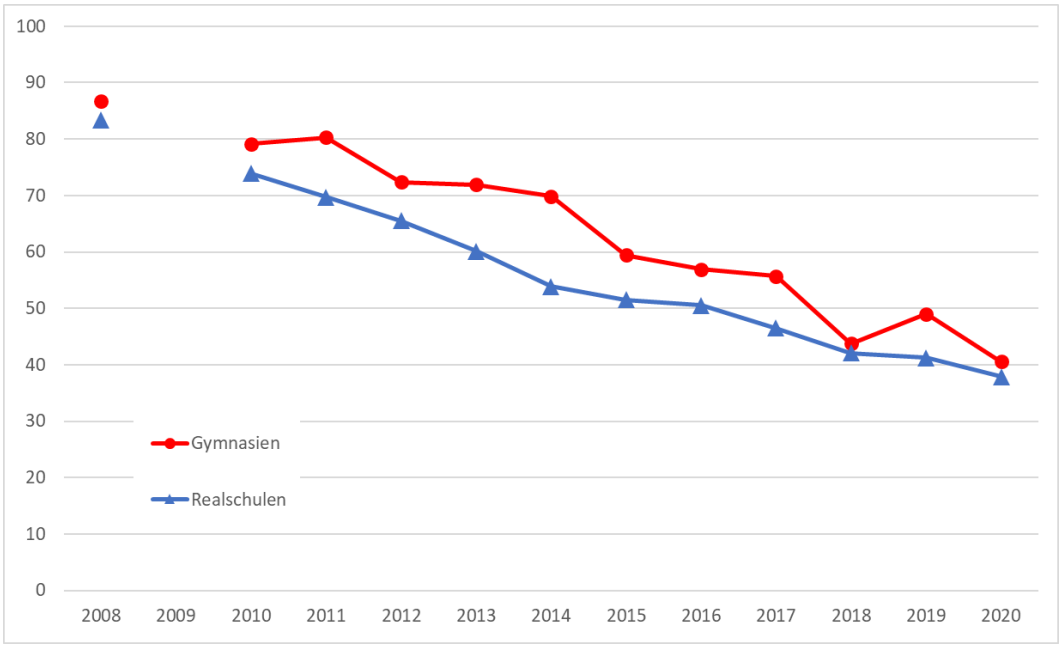


Abbildung 6.2.: Abruf von Rückmeldungen in Thüringen für weiterführende Schulen nach Schulform über die Jahre

direkter Zusammenhang bestehen. Zwei Veröffentlichungen berichten parallel dazu unterschiedliche Befunde bezüglich der Veränderung der Auseinandersetzung der Lehrkräfte mit den Rückmeldungen. So findet Groß Ophoff (2013, S. 297), dass sich die Intensität der Auseinandersetzung mit den Rückmeldungen im von ihr untersuchten Zeitraum von 2004 bis 2008 reduziert hat (mittlerer Effekt), so wie auch die direkte und indirekte Nutzung (kleiner Effekt). Parallel dazu hat sich die persönliche Wahrnehmung der Lehrkräfte, dass VERA Kontrollzwecken dient verstärkt und dass VERA das Ziel der Schul- und Unterrichtsentwicklung verfolgt verringert. Diese Beobachtungen für VERA in der Primarstufe decken sich mit den sinkenden Downloadszahlen der Rückmeldungen in Thüringen für die später in die Vergleichsarbeiten einsteigende Sekundarstufe. Groß Ophoff (2013) führt als Erklärung einen Neuigkeitseffekt an und beschreibt diesen als „kurzzeitige Motivations- bzw. Akzeptanzsteigerung [...], die bei zunehmender Vertrautheit mit dem System wieder zurückgeht“ (ebenda S. 297). Solche Neuigkeiten, die eine temporär verstärkte Nutzung begünstigen, beschreibt sie aber nicht allein für die Einführung von Vergleichsarbeiten, sondern auch für die Verschiebung des Termins in der Primarstufe von der Klassenstufe 4 auf 3. Demgegenüber findet Schliesing (2017) in ihrer Analyse von Rezeptionsstudien Hinweise für „eine verstärkte Auseinandersetzung mit VERA-Ergebnissen in Schulen und eine stärkere Nutzung dieser“ (ebenda S. S. 56) in Befunden der Jahre 2013 und 2014 gegenüber solchen aus den ersten VERA-Jahren.

Die über mehrere Jahre erhobenen Downloadzahlen in Thüringen geben zudem auch einen

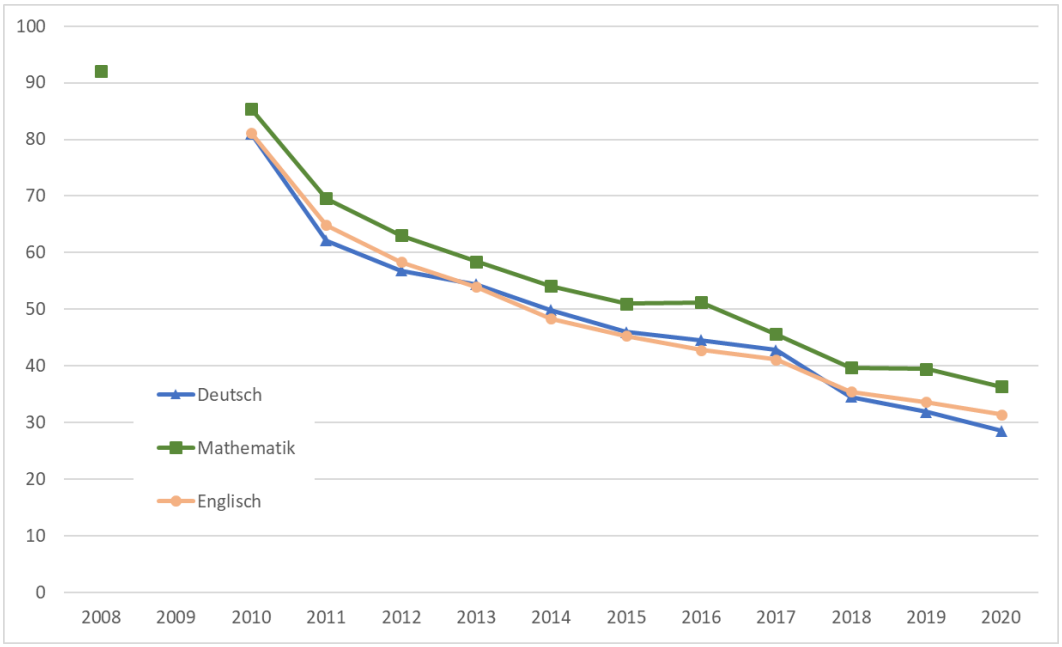


Abbildung 6.3.: Abruf von Rückmeldungen in Thüringen für VERA-8 und verschiedene Fächer über die Jahre

Hinweis darauf, ob der bundesweite Lockdown zur Eindämmung der Corona-Pandemie im März 2020 kurz nach der Durchführung von VERA-8 einen Einfluss auf die Abrufe hatte. Einerseits waren die Schulen damit konfrontiert neue Formen des Unterrichtens zu entwickeln, die sich nur auf eine eher unterentwickelte Digitalisierung der Schule stützen konnte und sich in vielfacher Hinsicht durch große Heterogenität bei der Ausstattung der Schulen, den Fähigkeiten der Lehrkräfte und den Möglichkeiten der Schülerinnen und Schüler auszeichnete. Andererseits entstanden auch Freiräume durch ausfallenden Unterricht und die teilweise aber fortgesetzte Präsenzpflicht der Lehrkräfte in den Schulen, die - so lässt sich zumindest anekdotisch berichten - auch für eine vertiefte Analyse der vorliegenden Vergleichsarbeiten und deren Rückmeldungen genutzt werden konnte. Die Abrufquoten Thüringens zeigen aber sowohl schulform- wie fachbezogen für 2020 eine unauffällige Fortsetzung des abfallenden Trends.

6.3. Forschungsfragen

Die vorliegende Untersuchung soll die Thüringer Perspektive auf das Thema der Rückmeldeabrufe ergänzen. Ohne theoretische Basis können hier keine Hypothesen formuliert werden. Dennoch sollen die Abrufe der Rückmeldungen bezüglich bestimmter Aspekte untersucht werden. Bei der Formulierung von Forschungsfragen wird auf die Thüringer Untersuchungen und

die Implementation von VERA in der Bildungsregion Berlin-Brandenburg Bezug genommen. Durch den Einbezug der Downloadzahlen für die beiden Länder Berlin und Brandenburg lassen sich zudem auch landesspezifische Besonderheiten betrachten und im Verhältnis zu den Ergebnissen aus Thüringen interpretieren.

Forschungsfrage 1: Unterscheiden sich die Abrufquoten von Rückmeldungen zwischen Gymnasien und nicht-gymnasialen Schulformen?

Die Abrufquoten Thüringens liegen für Gymnasien konsequent und im Mittel 7,5 Prozent über denen der Realschulen, allerdings sind hier wieder Rückmeldungen aus dem für alle Fächer verpflichtenden VERA-8 mit solchen aus dem nur für ein Fach verpflichtenden VERA-6 verknüpft. Da Gymnasien ihre Schülerinnen und Schüler in den Ländern Berlin und Brandenburg im Allgemeinen seit der Klassenstufe 7 mit dem Ziel der Erlangung des Abiturs ausbilden, stellen, so die Argumentation der Vereinigung der Oberstudiendirektoren des Landes Berlin e.V. (2017), die Prüfungen zum Mittleren Schulabschluss und damit auch die für diesen Zeitpunkt beschriebenen Bildungsstandards kein angemessenes Leistungskriterium für das Gymnasium dar. Aus diesem Grund wollen sich die Gymnasien auch der Pflicht entledigen, den Mittleren Schulabschluss verpflichtend abzunehmen. Von wissenschaftlicher Seite wurde dieser Forderung durch die Expertenkommission (Köller et al., 2020) entsprochen. Es kann also vermutet werden, dass die Vergleichsarbeiten von geringerem Interesse für die Lehrkräfte an Gymnasien sind als für solche anderer Schulformen. Vielleicht wird diesem Argument allerdings mit dem auf gymnasiale Ansprüche angepassten Testheft erfolgreich begegnet. Die relativ höheren Abrufquoten Thüringer Gymnasien könnten allerdings auch mit einer symbolischen Nutzung (Scheerens, 2007) von erwartbar guten Ergebnissen erklärt werden und so auch in Berlin und Brandenburg zu finden sein.

Forschungsfrage 2: Unterscheiden sich die Abrufquoten für Schulen aus den zwei Ländern?

Seit der Einführung der Vergleichsarbeiten sind diese insbesondere von Lehrkräften und der Gewerkschaft Erziehung und Wissenschaft (GEW) Berlins kritisiert worden. Im Zentrum der Kritik standen dabei das auf die zentralen Bildungsstandards zurückgehende für alle Schülerinnen und Schüler identische Schwierigkeitsniveau, auch für jene aus sozial benachteiligten Quartieren, wie sie sich häufiger in Berlin finden. Aber auch der zeitliche Aufwand der Beschäftigung mit den Tests allgemein wurde kritisiert (zum Beispiel GEW Berlin, 2013). Der

sich dabei nahezu ausschließlich in Berlin manifestierende Widerstand, ist inzwischen leiser geworden, könnte sich aber in sichtbar niedrigeren Abrufquoten zeigen. Weiter ist sicher interessant, inwieweit sich die Abrufquoten von denen Thüringens unterscheiden, wenngleich aktuell nur die Ergebnisse für einen einzigen Messzeitpunkt und auch nur für VERA-8 zur Verfügung stehen.

Forschungsfrage 3: Gibt es einen Zusammenhang zwischen der durchschnittlich erreichten Leistung und den Abrufquoten?

Bezüglich eines Zusammenhangs der Leistungsverteilung und der Ausschöpfung lassen sich zwei Vermutungen formulieren. Einerseits könnten eher Schulen im oberen Leistungsbereich ein Interesse daran haben, ihre Rückmeldungen im Sinne einer symbolischen Nutzung zu verwenden und dafür stärker auf diese zugreifen. Umgekehrt könnte eine Nutzung der Ergebnisse für Unterrichtsentwicklung für diese Schulen als weniger relevant angesehen werden, weshalb ein größeres Desinteresse vermutet werden kann. Äquivalente Argumente lassen sich für Schulen im unteren Leistungsbereich finden. Beide Effekte könnten sich aber ebenso ausmitteln.

6.4. Methode

6.4.1. Prozesse der Datenaquise und Analyse

Untersucht wurde die Durchführung von VERA-8 im Jahr 2020. Für zurückliegende Jahre liegen die für die Analyse notwendigen Daten nicht vor, so dass zum aktuellen Zeitpunkt keine Trendanalyse vorgenommen werden kann.

Berliner und Brandenburger Schulen nutzen für die Durchführung der Vergleichsarbeiten das durch das ISQ zur Administration bereitgestellte ISQ-Portal. Die Schulleitung legt hier mit ihrem Zugang fest, welche der Klassen und Kurse ihrer achten Jahrgangsstufe an VERA teilnehmen und übergibt dann ein projektspezifisches Passwort an alle beteiligten Lehrkräfte sowie ggf. an Fachkoordinator*innen. Die Lehrkräfte ergänzen wiederum Daten der Schülerinnen und Schüler und bereiten damit das Portal für die Eingaben der Ergebnisse vor. Die Tests finden in beiden Ländern an drei festgelegten Terminen statt. Schulen in privater Trägerschaft können an den Vergleichsarbeiten freiwillig teilnehmen, bis zu 50% nutzen dieses Angebot. Schülerinnen und Schüler an öffentlichen Schulen sind zur Teilnahme verpflichtet, sofern sie nach dem regulären Rahmenlehrplan unterrichtet werden. Ab 13 Uhr am jeweiligen Testtag können die Ergebnisse von den Lehrkräften eingegeben werden. Alle drei

Testtage lagen 2020 in einer Woche: Der Testtag für Deutsch war Montag der 2. März 2020, die Fremdsprachen folgten zwei Tage später am Mittwoch und Mathematik am Freitag dem 6. März. Die Eingabe der Ergebnisse eines Faches für eine Lerngruppe[3] wird durch die Lehrkraft abgeschlossen. Anschließend wird die Berechnung der Ergebnisse für diese Lerngruppe automatisch gestartet. Wird für alle Lerngruppen innerhalb eines Faches die Eingabe für die letzte Lerngruppe einer Schule abgeschlossen, werden durch einen weiteren automatischen Job die Ergebnisse für die gesamte Schule berechnet. Diese Berechnungen sind jeweils nach 3 bis 8 Minuten abgeschlossen und die erste Rückmeldung steht zur Verfügung.

- Die *erste klassenbezogene Rückmeldung* (auch *Sofort-Rückmeldung*) enthält die Lösungshäufigkeiten aller Aufgaben der einzelnen Testdomänen für die gesamte Lerngruppe. Da der Test aus Aufgaben unterschiedlicher Schwierigkeit besteht, werden die Lösungshäufigkeiten an jener gespiegelt, die das IQB im Rahmen der Pilotierung an der für Deutschland repräsentativen Erhebung festgestellt hat. Diese Rückmeldung wird für jede Lerngruppe und jedes Fach einzeln zur Verfügung gestellt.

Alle Rückmeldungen werden als PDF-Dateien zum Download bereitgestellt. Der Eingabezeitraum für die Ergebnisse endete offiziell am 27. März, das faktische Ende war ca. eine Woche später. Zu diesem Zeitpunkt sind üblich fast alle Ergebnisse der Schülerinnen und Schüler erfasst. Die Schließung der Schulen auf Grund der Corona-Pandemie führte zu einer Verzögerung der Abläufe nach den Testungen. Zwar war der Eingabestand für eine Berechnung von Vergleichswerten angemessen hoch, der Eingabezeitraum wurde trotzdem um drei Wochen verlängert. Damit auf der Basis eines definierten Datenbestandes Werte für Vergleichsgruppen über mehrere Schulen hinweg aggregiert werden können, wie beispielsweise Landeswerte, Werte für Schulen gleicher Schulform oder sozial adjustierte Vergleichswerte, wird das Portal dann vorübergehend geschlossen. Nach diesen Berechnungen und einer Plausibilitätsprüfung öffnet das Portal wieder, so dass ggf. fehlende Daten nachgetragen werden können. Der Anteil derart später nachgetragener Ergebnisse liegt aber deutlich unter einem Prozent, kann also für Analysen dieser Untersuchung vernachlässigt werden. Nach der Berechnung der verschiedenen Vergleichswerte werden sämtliche weitere Rückmeldungen freigeschaltet. 2020 erfolgte diese Freischaltung aller anderen Rückmeldungen gleichzeitig am 4. Mai. Ohne pandemiebedingte Verzögerung der Dateneingabe, würde diese Freischaltung sukzessive geschehen, wobei aber auch dann bis zum 4. Mai sämtlicher Rückmeldungen freigeschaltet worden wären. Jede

[3] Hier wird der Begriff Lerngruppe verwendet, weil insbesondere die Vergleichsarbeiten in der Sekundarstufe I teilweise in Klassen und teilweise in Kursen absolviert werden.

dieser Rückmeldungen enthält die Ergebnisse zu den Testdomänen eines Faches. Es sind im Einzelnen:

- Die *individuelle Rückmeldung* enthält für jede teilnehmende Schülerin und jeden teilnehmenden Schüler eine Doppelseite mit den Ergebnissen des VERA-Test für ein Fach. Hier findet sich die durchschnittliche Lösungshäufigkeit aller Aufgaben einer jeden Domäne, welcher die durchschnittliche Lösungshäufigkeit der Lerngruppe gegenübergestellt ist. Zudem wird die Verteilung der Schülerinnen und Schüler der Lerngruppe auf die Kompetenzstufen berichtet und das individuelle Ergebnis eingeordnet. Seit mehreren Jahren wird diese individuelle Rückmeldung durch ein Konfidenzintervall ergänzt, um Bedenken bezüglich der mangelhaften Präzision dieser Zuordnung Rechnung zu tragen.

- Die Vergleichsarbeiten zielen im Kern auf eine Nutzung der Ergebnisse zur Analyse der Unterrichtsergebnisse der Schülerinnen und Schüler, um daraus Schlussfolgerungen für eine zielführende Weiterarbeit abzuleiten. Im Zentrum der Rückmeldungen steht deshalb die zweite *klassenbezogene Rückmeldung*, welche die Ergebnisse detailreich beleuchtet. Konkret werden hier die Lösungshäufigkeiten der Lerngruppe für jede Domäne und auch für einzelne Teilkompetenzen ausgewiesen und denen der Schule und einer Vergleichsgruppe gegenübergestellt. Teil der Rückmeldung ist ebenso die graphische Darstellung der Verteilung der durch die Schüler*innen erreichten Kompetenzstufen, sowie ein tabellarischer Überblick über die Ergebnisse der einzelnen Schülerinnen und Schüler. Auch diese Rückmeldung wird für jedes einzelne Fach erstellt.

- In der *schulbezogenen Rückmeldung* werden die Kompetenzstufenverteilungen aller Lerngruppen jener der eigenen Schule gegenübergestellt sowie der von Schulen der gleichen Schulform und Schulen ähnlicher sozialer Struktur. Die fachbezogene Rückmeldung weist zudem die einzelnen getesteten Domänen aus.

Im Anhang A.7.2 sind für jede Rückmeldung Beispiele abgebildet, wie sie für das Land Berlin durch das ISQ zur Verfügung gestellt werden. Die Rückmeldungen für Brandenburg unterscheiden sich nur marginal[4].

6.4.2. Daten

Alle Rückmeldungen werden als PDF-Dateien angeboten, die von Lehrkräften, Fachkoordinator*innen und Schulleitungen aus dem Portal heruntergeladen werden können. Dabei haben

[4] Bisher haben sich für eine sozial adjustierte Rückmeldung an Brandenburger Schulen keine angemessenen Parameter finden können, weshalb auf solche Werte in der Rückmeldung für Brandenburg verzichtet wird.

Tabelle 6.1.: Erste und mehrfache Abrufe verschiedener Rückmeldungen im Rahmen von VERA-8 2020 für Berlin und Brandenburg

Art der Rückmeldung	Abrufe	davon erste Abrufe (Anzahl)	(relativ)	durchschnittliche Abrufhäufigkeit
Sofort	14.088	4.713	33,5%	3,0
Individual	6.784	3.695	54,5%	1,8
Klasse	5.632	3.353	59,5%	1,7
Schule	1.841	824	44,8%	2,2
Summe	28.345	12.585	44,4%	2,3

alle Rezipienten Zugriff auf sämtliche Rückmeldungen. Im hier betrachteten Zeitraum vom 7. Januar bis zum Ende des 25. Oktober 2020 erfolgten 37.072 Downloads, von denen 3.018 im Rahmen der Softwareentwicklung, vom ISQ-Projektmanagement zur Prüfung der Funktionalität oder von der Hotline ausgeführt wurden. Die verbleibenden 34.054 Downloads entfallen mit 72,0% zum großen Anteil auf Lehrkräfte, weitere 23,6% auf Schulleitungen und der verbleibende Rest auf Fachkoordinator*innen (4,4%). Reduziert man diese Downloads auf jene, die dem aktuellen Projekt VERA-8 2020 zuzurechnen sind, verbleiben für die weitere Analyse 30.648 Download-Aktivitäten, von denen aber lediglich 28.345 aus öffentlichen Schulen ausgewertet werden sollen. Die folgenden Analysen sollen zudem nur solche Abrufe fokussieren, bei denen eine Rückmeldung das erste Mal abgerufen wurde (vergleiche Tabelle 6.1). Ein erfolgter erster Abruf einer Rückmeldung ist die Mindestvoraussetzung dafür, dass die Ergebnisse der Vergleichsarbeiten genutzt werden können. Die Rückmeldung kann erst dann in der Schule verfügbar sein. Jede Sofortrückmeldung wird durchschnittlich 3 Mal aufgerufen, Schulrückmeldungen etwas mehr als 2 Mal, Individual- und Klassenrückmeldungen etwas weniger als 2 Mal. In die Analyse gehen demnach 12.585 erste Abrufe ein.

Die Zahl der Downloads wurde auf die Zahl der absolvierten und ins Portal eingegebenen Tests und damit auf die Zahl der vorgehaltenen Rückmeldungen bezogen und nicht auf alle angemeldeten Schulen bzw. Lerngruppen. Dass im Land Brandenburg jeweils eine der zwei Domänen in den Fächern Deutsch und Englisch nur optional durchgeführt werden musste, kann für die folgende Betrachtung unberücksichtigt bleiben, weil sich die Aussagen zu den Ergebnisdownloads immer auf das gesamte Fach beziehen.

6.4.3. Operationalisierung

Schon oben wurde diskutiert, wie die mit den Abrufen operationalisierte Variable zu interpretieren ist. Die tatsächlichen Prozesse der Verwertung der Testergebnisse bleiben auch nach

dieser Untersuchung weiter im Verborgenen. Selbst die vorliegenden Daten der Abrufe von Rückmeldungen lassen Raum für unterschiedliche Verfahrensweisen. So ist vorstellbar, dass die Schulleitung oder eine Fachkonferenzleitung alle Rückmeldungen aus dem Portal herunterlädt, um Auswertungsprozesse gleich oder später anzustoßen. Unklar ist auch, was ein mehrfacher Abruf einer Rückmeldung bedeutet. Zudem ist unsicher, wer eine Rückmeldung abruft. Das Management der Nutzer*innen im ISQ-Portal erlaubt die Unterscheidung von Schulleitung[5], Lehrkräften und Fachkoordinator*innen, kann aber einzelne Personen nicht identifizieren und auch keiner konkreten Lerngruppe zuordnen.

Als Sicher kann lediglich gelten, dass Ergebnisse einer nicht abgerufenen Rückmeldung in der Schule nicht verwertet werden können, zumindest, wenn der betrachtete Zeitraum ausreichend groß gewählt wird.

6.4.4. Analysestrategie

Im Folgenden werden für jede Rückmeldeart die Zahl sämtlicher ersten Abrufe von allen in den Schulen zum Zugriff Berechtigten jener Zahl von Rückmeldungen gegenübergestellt, die zum Abruf erzeugt und vorgehalten wurden. Die absoluten Zahlen sind demnach als Vollerhebung für die Abrufe der Schulen Berlins und Brandenburgs zu interpretieren.

In der beispielhaften Darstellung einer Sofort-Rückmeldung in Abbildung 6.4 werden die Abrufzeitpunkte auf der horizontalen Zeitachse durch kleine schwarze Striche dargestellt. Die Abrufe liegen insbesondere in den ersten Tagen und Wochen nach der Freischaltung im März so dicht, dass sie in dieser Darstellung nicht mehr einzeln identifiziert werden können, was eine vermutete Häufigkeit früher Abrufe bestätigt. Diese Rückmeldeabrufe werden für die folgenden Analysen tageweise aggregiert. In der Graphik sind die aggregierten Werte durch schmale blaue Balken wiedergegeben. Für die Analyse ist letztendlich die Summe der abgerufenen Rückmeldungen von besonderem Interesse, konkret als prozentualer Anteil der abgerufenen von den zur Verfügung gestellten Rückmeldungen. Eher sekundär ist die zeitliche Verteilung der Abrufe. Deshalb wurde zur Visualisierung eine kumulative Darstellung der relativen Rückmeldeabrufe gewählt (rote Kurve in der Abbildung 6.4). Eine äquivalente Darstellung findet sich auch in den Thüringer Landesberichten (jüngst Nachtigall, 2020).

[5] Auch das Passwort der Schulleitung ist nicht eindeutig einer Person zuzuordnen. In den Bedingungen des Umgangs mit dem Passwort ist es der Schulleitung erlaubt, das Passwort mit einer weiteren Person zu teilen. So soll sichergestellt werden, dass die Nutzung des ISQ-Portals im Fall der Verhinderung der Schulleitung möglich bleibt. Zudem wurde deutlich, dass einige organisatorische, der Schulleitung allein zufallende Aufgaben im Rahmen der Initiierung von Projekten im Portal, durch diese aus zeitlichen Gründen schwer allein bewältigt werden können. Durch die klare Regelung wurde einer missbräuchlichen Nutzung des Zugangs der Schulleitungen entgegengewirkt.

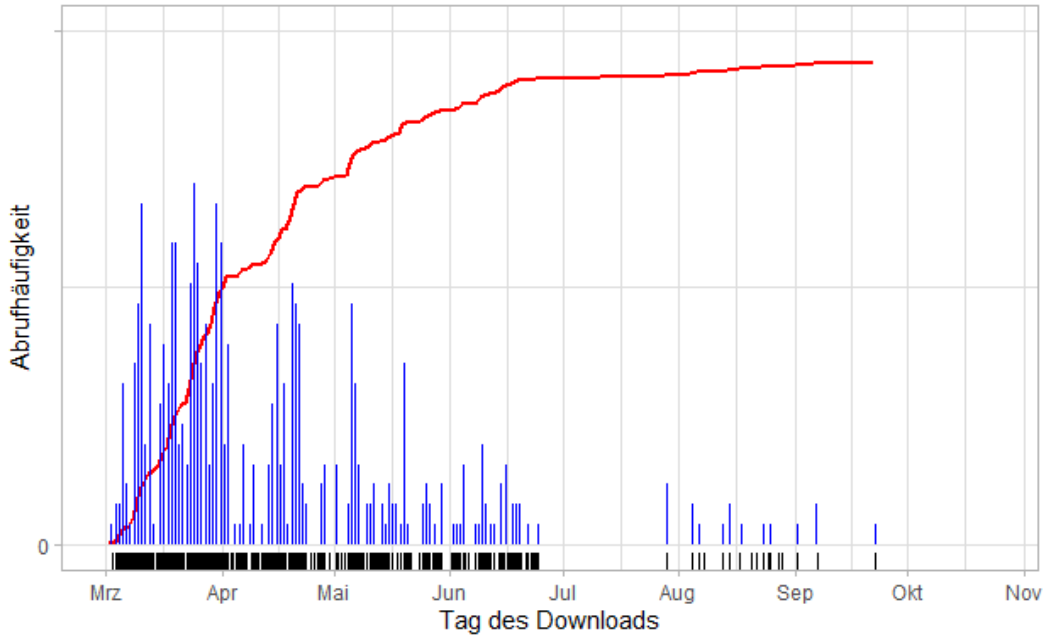


Abbildung 6.4.: Beispielhafte Darstellung von Rückmelde-Downloads über die Zeit

Der letzte Wert der monoton steigenden *Abrufquote* im untersuchten Zeitraum stellt dann die finale Abrufquote dar, die hier als *Ausschöpfung* bezeichnet wird. Sie kann als endgültig angesehen werden, wenn die Abrufe bzw. die Steigung der kumulativen Funktion der Abrufe zum Ende des Untersuchungszeitraumes nahe Null ist. Als Parameter dieser Funktion werden des Weiteren die Abrufquote zu einem bestimmten Zeitpunkt nach Freischaltung sowie die Zeit bis zum Erreichen einer bestimmten Abrufquote analysiert.

Der Verlauf der Funktion der kumulativen Abrufe A_k über die Zeit t seit der Freischaltung der Rückmeldung bei $t = 0$ kann mit der Funktion

$$A_k = a_{asy}(1 - e^{-st}) \tag{6.1}$$

mathematisch modelliert werden. Diese Form wird üblich für die Modellierung von Wachstumsprozessen verwendet. Negative Exponenten verweisen dabei auf ein beschränktes Wachstum. Im vorliegenden Fall gibt es mit der Anzahl der zur Verfügung stehenden Rückmeldungen sogar eine natürliche, angebbare Schranke für das Wachstum. In der Funktion beschreibt der Parameter s die Steigung der Funktion. Der zweite Parameter a_{asy} steht für die *asymptotische Ausschöpfung*, also jener Wert der kumulativen Abrufe, dem die Funktion im unendlichen zustrebt. Dieser Wert bildet also die obere Grenze dessen, was dem Model nach für eine bestimmte Rückmeldung maximal erwartet werden kann. Geht man davon aus, dass alle Rückmeldungen abgerufen werden, liegt dieser Wert bei 100%.

Tabelle 6.2.: Beispiel für das Ergebnis einer nichtlinear modellierten Abruffunktion

Art		nichtlineare Regression				Anzahl Tage[b]		Ausschöpfung[c]	
	Para[a]	est	st.err	t	sig	abs	[%]	[%]	CI(95%)
Sofort	a	0.9711	0.0039	164.35	0.000	120	50.6	97.1	0.36
	s	-0.0222	0.0003	-63.88	0.000				

[a]Parameter der Regression: a = finale Ausschöpfung, s = Steigung der Funktion
[b]Anzahl der Tage, an denen mindestens ein Abruf erfolgte.
[c]geschätzte Ausschöpfung für das Ende des Untersuchungszeitraumes, 25.10.2020.

Mit Hilfe einer nichtlinearen Regression wird für den Untersuchungszeitraum untersucht, ob das Modell Gültigkeit besitzt. Zudem lassen sich messbedingte Streuungen über ein Konfidenzintervall abbilden und ggf. zwei Abruffunktionen bezüglich der geschätzten Ausschöpfung als überzufällig unterschiedlich identifizieren. Ob diese Modellierung über den hier betrachteten Zeitraum von 8 Monaten hinaus bestand hat, kann nur in einer späteren Untersuchung gezeigt werden. Für die vorliegenden Betrachtungen ist dies allerdings von untergeordnetem Interesse. Die Tabelle 6.2 zeigt ein Beispiel dafür, wie das Ergebnis der Modellierung berichtet wird. Zusätzlich wird die Zahl der Tage angegeben, an denen überhaupt Rückmeldungen abgerufen wurden und zur Gesamtzahl der Tage im Untersuchungszeitraum ins Verhältnis gesetzt. Die Ausschöpfung entspricht dem Parameter a und wird durch die Breite des abgeleiteten Konfidenzintervalls ergänzt.

Die gewählte Modellierung muss als trivial bezeichnet werden. Mathematisch, weil sie mit nur zwei Parametern versucht sämtliche Abrufverlaufsfunktionen zu modellieren und inhaltlich, weil sie keine erwartbaren Einflüsse einbezieht, wie zum Beispiel eine Abflachung der kumulativen Abrufe über die Sommerferien oder wieder zunehmende Abrufe zum Schuljahresanfang. Trotzdem wird erwartet, dass diese Modellierung für einen statistisch abgesicherten Vergleich von Abruffunktionen im Rahmen dieser ersten explorativen Untersuchung ausreichend sein wird.

Die Forschungsfragen (Abschnitt 6.3) erfordern die Untersuchung des Einflusses von verschiedenen Prädiktoren auf die binäre Variable, ob eine Rückmeldung letztendlich abgerufen wurde oder nicht. Das Verfahren der logistischen Regression erlaubt solche Untersuchungen. Die in den Annahmen beschriebenen, den Rückmeldeabruf beeinflussenden Parameter werden dabei als unabhängige Prädiktoren berücksichtigt. Beim Einbezug mehrerer Prädiktoren sollten wie bei einer linearen Regression auch, Kollinearitäten ausgeschlossen werden. Die Bewertung der Ergebnisse einer logistischen Regression wird durch zwei Umstände erschwert. Die unabhängigen Variablen werden in der Modellfunktion als Exponenten einer e-Funktion

formuliert. Dadurch sind die Parameter nicht direkt bewertbar. Zudem sind die Zusammenhänge nichtlinear. Das bedeutet, dass die Wirkung einer unabhängigen Variablen in verschiedenen Skalenbereichen unterschiedlich ist. Als Ergebnis einer linearen Regression formuliert man beispielsweise: Eine Änderung des unabhängigen Parameters X um eine Einheit bewirkt eine Veränderung des abhängigen Parameters Y um z. Dies gilt in dieser Form für die Parameter einer logistischen Regression nur für eine bestimmte Ausprägung von $X = X_1$. Um das Verständnis für die Regression mit mehreren Parametern vorzubereiten, wird im Folgenden als Vorgriff auf die Untersuchung der Zusammenhänge im Abschnitt 6.5.2 für die erste Forschungsfrage und bezogen auf die Schulrückmeldung eine beispielhafte logistische Regression mit einer unabhängigen Variable berechnet und interpretiert.

In den folgenden Darstellungen ist immer zu unterscheiden, ob der zeitliche Verlauf der Abruffunktion mit Hilfe einer nichtlinearen Regression modelliert wird oder ob der Einfluss einer Variable auf den endgültigen Abruf mit Hilfe der logistischen Regression bestimmt wird.

Beispielhafte logistische Regression

Die abhängige Variable, ob ein Abruf der entsprechenden Rückmeldung vorliegt (1) oder ob die Rückmeldung nicht abgerufen wurde (0), wird für die Berechnung entsprechend binär kodiert. Zur leichteren Interpretation der Ergebnisse wurde die Schulform als Unabhängige ähnlich kodiert: 0 = nicht gymnasiale Schulformen und 1 = Gymnasien. Das Basismodell, bei dem die unabhängige Variable gleich Null gesetzt wird, beschreibt also die Wahrscheinlichkeit des Abrufs einer Rückmeldung für alle Schulen, die keine Gymnasien sind.

Ausgangsbasis des Beispiels ist die rot eingefärbte 2x2-Tafel im die Schulrückmeldung beschreibenden oberen Teil der Tabelle 6.3. Für jede Ausprägung der hier untersuchten unabhängigen Variable *Schulform* wird die Wahrscheinlichkeit für einen erfolgten Abruf $P(Abruf = 1)$ berechnet. Der sogenannte Odd stellt diese Größe in anderer Form als Quotient aus Wahrscheinlichkeit und Gegenwahrscheinlichkeit dar, ein Verhältnis das auch als Chance oder Wettquotient bezeichnet wird. Für die nicht gymnasialen Schulformen, welche die Rückmeldungen zu etwas mehr als 50% abrufen, ergibt sich damit

$$Odds_{sf} = \frac{\frac{143}{279}}{\frac{136}{279}} = \frac{143}{136} = 1,0515$$

für die Gymnasien berechnet man

$$Odds_{sf} = \frac{\frac{116}{167}}{\frac{51}{167}} = \frac{116}{51} = 2,2745$$

Die Analyse mit Hilfe der logistischen Regression gibt hier letztendlich das Verhältnis beider Quotienten, dass sogenannte Odds Ratio (OR) zurück

$$OR = \frac{2,2745}{1,0515} == 2,1631$$

Im vorliegenden Fall würde also geschlussfolgert: Die *Chance*, dass eine Rückmeldung abgerufen wird, ist im Gymnasium etwa 2,16 Mal höher als an anderen Schulen. Dies bedeutet nicht, dass doppelt so viele Rückmeldungen abgerufen werden; vorliegend ist das Verhältnis ca. 50 zu 70%. So wie Odds als Chance oder Quote bezeichnet und interpretiert werden, wird das OddsRatio als Chancen- oder Quotenverhältnis oder auch als Wettquotient bezeichnet. Diese Bedeutungen helfen bei inhaltlichen Interpretationen allerdings selten. Zur Interpretation der Modellergebnisse von logistischen Regressionen stellen denn auch Best und Wolf (2012) fest, dass „die meisten Menschen [...] nicht in der Lage sind, die Bedeutung von OR in Bezug auf ein substanzielles Problem richtig zu erfassen“ (ebenda S.381). Dies gilt insbesondere, wenn die Basiswahrscheinlichkeiten weit weg von 50% liegen. Eine sinnvolle Interpretation von OddsRatio ist überhaupt nur bei bekannten Basiswahrscheinlichkeiten möglich. Anders, als bei üblichen Problemen, sind diese hier tatsächlich bekannt. Für eine Verdeutlichung des Problems, wird die Regressionsgleichung für die schulbezogene Rückmeldung im Fach Mathematik konkret formuliert (vergleiche Ergebnis der Regression in Tabelle 6.4):

$$P(Abruf = 1) = \frac{1}{1 + e^{-(0,0502+0,7716 \cdot Schulform)}}$$

Für $Schulform = 0$ (Basismodell der nicht gymnasialen Schulen) ergibt sich die Wahrscheinlichkeit für einen Abruf zu

$$P(Abruf = 1) = \frac{1}{1 + e^{-(0,0502)}} = 0,5125$$

was der realen Basiswahrscheinlichkeit (siehe Tabelle 6.3) exakt entspricht, so wie die auf der Basis des Regressionsmodells berechneten Wahrscheinlichkeit für Gymnasien mit $Schulform = 1$

$$P(Abruf = 1) = \frac{1}{1 + e^{-(0,0502+0,7716 \cdot 1)}} = 0,6946$$

Dieses logistische Regressionsmodell kann die realen Verhältnisse für den vorliegenden Fall demnach sehr gut beschreiben.

Führt man sich die Modellgleichung vor Augen, werden die Schwierigkeiten der Interpre-

Tabelle 6.3.: Einfluss der Schulform auf die Häufigkeit der Abrufe, beispielhaft für die schulbezogene und die Sofort-Rückmeldung im Fach Mathematik

Art	Schulform	Abrufe 0	Abrufe 1	$\sum$	$P(Abruf = 1)$	$Odds_{sf}$	$OddsRatio$	$ln(OR)$
Schule	0 (nGy)	136	143	279	0.5125	1.0515	2.1631	0.7715
	1 (Gy)	51	116	167	0.6946	2.2745		
Sofort	0 (nGy)	715	579	1294	0.4474	0.8096	19.3667	2.9636
	1 (Gy)	41	643	684	0.9401	15.6829		

tation, insbesondere im Vergleich zu einer linearen Regression, offensichtlich: Zuerst liegt die Bedeutung des Intercepts von -0,0502 in der Abweichung der Basiswahrscheinlichkeit von 50%. Wie schon oben ausgeführt ist dies die Basiswahrscheinlichkeit für nicht gymnasiale Schulen, weil für diese die Schulform im Datensatz mit 0 kodiert ist, womit der $ln(OR)$-Koeffizienten[6] von 0,7716 für die Schulform mit Null multipliziert wird. Der Intercept liegt im Beispiel nahe Null, so dass sich die Wahrscheinlichkeit wegen $e^0 = 1$ eben zu gut 50% ergibt. Das entsprechende Odd liegt in einem solchen Fall bei 1. Schon für die Gymnasien liegt die Wahrscheinlichkeit bei knapp 70% und das Odd bei über 2. Hier wird die Nichtlinearität dieser Beziehung deutlich: Für die Sofort-Rückmeldung an Gymnasien (siehe Abbildung 6.3, unterer Teil) bei 94% wird das Odd rund 16. Ist das Odd als Quotient von Wahrscheinlichkeit und Gegenwahrscheinlichkeit wegen des nichtlinearen Zusammenhangs schon schwer zu interpretieren, so ist die Bedeutung des Verhältnisses zweier Odds kaum mehr auf das Problem bezogen zu bewerten.

Deshalb wird die Interpretation des Ergebnisses von logistischen Regressionsanalysen (siehe Tabelle 6.4) oft wie auch in der vorliegenden Arbeit darauf beschränkt (a) die Bedeutung des Effekts eines Parameters als signifikant oder nicht signifikant zu klassifizieren sowie (b) die Richtung des Effekts anzugeben. Zudem kann mit R^2 und Cohens d der Effekt der im Modell einbezogenen unabhängigen Variablen bestimmt und mit der Deviance bzw. dem AIC die Modellpassung quantifiziert und ggf. zwischen verschiedenen Modellen verglichen werden. Da die Basiswahrscheinlichkeiten einfach abzuleiten bzw. in den Abruf-Graphiken gut visualisiert sind, können die Interpretationshilfen von Best und Wolf (2012) in Form speziell korrigierter Parameter unbeachtet bleiben. Die Tabelle 6.4 zeigt die Ergebnisse der logistischen Regression für die Schulrückmeldungen im Fach Mathematik. Das Odds Ratio ergibt sich als Potenz $e^{est} = e^{0,7716} = 2,1632$ des durch die Regression geschätzten Parameters. Überstreicht das in der letzten Spalte angegebene Konfidenzintervall CI(95%) für diesen Parameter die 1, so

[6] In der allgemeinen Formulierung wird dieser Koeffizient oft als *beta*-Koeffizient bezeichnet.

Tabelle 6.4.: Beispielhafte Ergebnisse einer logistischen Regression

Art[a]	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Schule	(Intercept)	0.0502	0.1198	0.4190	0.6752	1.0515	[0.8314; 1.3304]
	Schulform	0.7716	0.2063	3.7394	0.0002	2.1632	[1.4495; 3.2581]
	R^2	Cohens d			Deviance	dof	AIC
	0,0430	0,2119			592,13	444	596,13

[a]Die Bezeichnung *Art* meint hier und in der Folge immer *Art der Rückmeldung.*

muss die Vermutung einer signifikant höheren ($e^{est} > 1$) oder geringeren Chance ($e^{est} < 1$) verworfen werden. Ob der Effekt als statistisch signifikant gilt, zeigt auch $Pr(> |z|) < 0.05$ an. Im Beispiel wird die 1 vom Intervall [1,4495; 3,2581] nicht eingeschlossen und die Abrufe an Gymnasien sind signifikant häufiger[7]. Für das gesamte Modell wird zudem ein R^2 angegeben, hier als Nagelkerke R-Quadrat ausgeführt. Es ist wegen der modellbildenden logistischen Funktion nicht exakt wie das Äquivalent einer linearen Regression als der Anteil aufgeklärter Varianz zu interpretieren. Höhere Werte des zwischen 0 und 1 liegenden R-Quadrat zeigen aber auch hier eine bessere Passung des Modells an die Daten an. Für eine Bewertung der Effektstärke wurde zudem Cohens d ausgewiesen und weitere Parameter für die Modellgüte, die bei konkurrierenden Modellen für eine Entscheidung herangezogen werden können.

In den Gymnasien wird die Schulrückmeldung in knapp 70% der Fälle aufgerufenen, in den anderen Schulformen mit etwa 50% hingegen seltener. Der Einfluss der Schulform auf den Abruf der Sofortrückmeldung ergibt sich mit $Pr(> |z|) = 0.0002$ als hoch signifikant, wobei der Effekt nach Cohen (1988)[8] als klein eingestuft wird.

6.5. Ergebnisse

Der Untersuchung der Forschungsfragen im Abschnitt 6.5.2 ist eine umfassende Deskription der Daten im Abschnitt 6.5.1 vorangestellt. Die Ergebnisdarstellung soll Antworten dazu geben, wie viele der zur Verfügung stehenden vier unterschiedlichen Ergebnisrückmeldungen in welchem Zeitraum für die einzelnen Fächer abgerufen worden sind. Eine entsprechende graphische Darstellung soll zudem zeigen, ob die Schulen Berlins und Brandenburgs untereinander und auch im Vergleich zu den Schulen Thüringens ein ähnliches Downloadverhalten

[7]Der Abruf ist mit 1 gegenüber 0 kodiert und die Schulform Gymnasium ebenso mit 1 gegenüber der 0 für nicht gymnasiale Schulformen. Deshalb muss geschlussfolgert werden, dass an Gymnasien mehr Abrufe erfolgen.

[8]Cohen beschreibt einen Effekt als klein, wenn d zwischen 0,2 und 0,5 liegt, als mittel, wenn d zwischen 0,5 und 0,8 liegt und bei größerem d als groß.

zeigen.

Die Abrufe verteilen sich erwartet unregelmäßig über die Tageszeiten, wobei in den 12 Stunden von 6 bis 18 Uhr fast 80% aller Abrufe stattfinden, fast 20% in den 6 Stunden bis Mitternacht und noch 1,5% in der Zeit nach Mitternacht. Dabei lassen sich weder für die Art der Rückmeldung noch für das Fach Auffälligkeiten feststellen.

6.5.1. Deskription der Daten

In je einem Abschnitt für jedes Fach werden in zwei Tabellen und einer Graphik die Daten präsentiert. Die jeweils ersten Tabellen (6.5, 6.7 sowie 6.9[9]) stellen die Ausschöpfung und ausgewählte Parameter der Abrufe gegenüber. Die Spalten enthalten im Einzelnen:

- Die Differenzierung zwischen den vier Arten von Rückmeldungen: (1) aufgabenbasierte *Sofort*rückmeldung, (2) *Indivi*dualrückmeludung für Schülerinnen und Schüler, (3) zentrale *klasse*nbasierte Rückmeldung und (4) die auf die *Schule* bezogene Rückmeldung (siehe auch Seite 177).

- In der zweiten Spalte steht die Anzahl der verfügbaren Rückmeldungen. Für jede von Lehrkräften abgeschlossene Lerngruppe werden die drei klassenbezogenen Rückmeldungen erzeugt, die sich auf die Ergebnisse eben dieser Lerngruppe beziehen. Diese Zahl ist deshalb für diese drei Rückmeldungen immer identisch. Sind die Ergebnisse aller Lerngruppen einer Schule für ein Fach abgeschlossen, wird automatisch eine schulbezogene Rückmeldung für dieses Fach erstellt. Werden die Fachtests in einer Schule nicht für sämtliche Lerngruppen abgeschlossen, wird auch keine schulbezogene Rückmeldung erstellt.

- Nur die Sofortrückmeldung wird wenige Minuten nach erfolgter Eingabe automatisch freigeschaltet, deshalb gibt es die Angabe, welcher Anteil der Rückmeldungen bis zum Ende des Eingabezeitraums abgerufen wurde, nur für diese Rückmeldung.

- In der Spalte „von Ausschöpfung 50% erreicht“ ist das Datum (bzw. die Anzahl der Tage) angegeben, bis zu dem 50% jener Lehrkräfte ihre Rückmeldungen abgerufen haben, die diese im Untersuchungszeitraum überhaupt abgerufen haben (letzte Spalten). Das Datum bezieht sich also *nicht* auf den Abruf von 50% aller vorgehaltenen Rückmeldungen (zweite Spalte). Daneben gibt es die äquivalenten Angaben für den Anteil von 90%.

[9] Für das Fach Französisch liegen nur wenige Daten vor, so dass diese erste Tabelle nicht berichtet wird.

- Die Anzahl und der relative Anteil der abgerufenen bezogen auf alle zur Verfügung gestellten Rückmeldungen (Ausschöpfung) ergänzt die Daten.

Der Untersuchungszeitraum überstreicht für die Sofortrückmeldung 237 Tagen (7,7 Monate) und für die anderen, später freigeschalteten drei Rückmeldungen 174 Tagen (5,7 Monate). In den darauf folgenden Graphiken (6.5, 6.6, 6.7 sowie 6.8) werden die Abrufe der vier Rückmeldearten über der Zeit kumulativ dargestellt. 100% entsprechen dabei einem Abruf aller zur Verfügung gestellten Rückmeldungen.

In den Tabellen unter der Graphik (6.6, 6.8, 6.10 sowie 6.11) wird die Ausschöpfung über das nichtlineare Regressionsmodell entsprechend der Formel 6.1 geschätzt. Die Tabellen weisen die Schätzungen für die zwei Parameter aus, die absolute und relative Anzahl der Tage, für die Downloads berichtet wurden sowie ein sich aus der Regression für die geschätzte Ausschöpfung ergebendes Konfidenzintervall. Im Allgemeinen sollte die Ausschöpfung zum Ende des Untersuchungszeitraumes aus der ersten Tabelle sehr ähnlich der durch die Regression geschätzten Ausschöpfung sein. Unterschiede deuten darauf hin, dass die Modellierung für diesen letzten Zeitpunkt nicht optimal passt. So ergibt sich aus der Regression für die schulbezogene Rückmeldung eine geschätzte Ausschöpfung von 54,7%. Tatsächlich sieht man am Verlauf der Kurve, dass zum Ende der Sommerferien vermutlich einige Schulleitung vermehrt einen Blick auf diese Rückmeldung werfen, so dass tatsächlich eine Ausschöpfung von über 58% erreicht wird. Umgekehrt ist die Abflachung der Kurve für die Abrufe der Sofortrückmeldungen mit dem Beginn der Sommerferien deutlich, so dass die geschätzte Ausschöpfung von 64,4% die tatsächliche von nur 61,8% leicht überschätzt. Die Regression bezieht solche Variationen gerade nicht ein. Wie schnell Rückmeldungen abgerufen werden, kann man den mittleren Spalten der oberen Tabelle mit der deskriptiven Abrufstatistik („von Ausschöpfung % erreicht“) entnehmen oder auch dem geschätzten Steigungsparameter s der Regression. Ein absolut kleiner Wert von s steht für eine geringere Steigung. Die Schulrückmeldungen werden offensichtlich deutlich schneller heruntergeladen, als andere. Dass die Werte allesamt negativ sind ist der Art der Modellierung zuzuschreiben und für die Interpretation ohne Belang.

Für das Fach Französisch liegen nur für 25 Lerngruppen aus 17 Schulen Ergebnisse vor. Französisch kann nur in Berlin als erste Fremdsprache gewählt werden. Die Rückmeldungen wurden bis zum Ende des untersuchten Zeitraumes für 10 der 17 Schulen und für 15 bis 19 der 25 Lerngruppen abgerufen. Wegen der geringen Zahl werden in der Folge keine detaillierten Parameter dargestellt und auch aus den folgenden Analysen wird Französisch ausgenommen. Die Graphik lässt aber erkennen, dass die Abrufhäufigkeit und die Zeiträume

Tabelle 6.5.: Downloads für das Fach Mathematik, deskriptiv

Art	Rückmeldungen verfügbar [Anzahl]	Ende der Eingabe[a] [in %]	von Ausschöpfung[b] 50% erreicht [Datum]	[Tage]	90% erreicht [Datum]	[Tage]	Ausschöpfung[c] [Anzahl]	[in %]
Sofort	1.978	34,9	30.03.	28	26.05.	85	1.222	61,8
Indivi	1.978	-	19.05.	15	21.06.	48	1.003	50,7
Klasse	1.978	-	18.05.	14	23.06.	50	903	45,7
Schule	446	-	13.05.	9	17.06.	44	259	58,1

[a]Die Abrufquote der Sofortrückmeldungen zum Ende der regulären Eingabezeit nach 34 Tagen.
[b]Zeitpunkt, zu dem 50% bzw. 90% der Ausschöpfung nach Freischaltung erreicht werden.
[c]Ausschöpfung nach 237 (Sofortrückmeldung) bzw. 174 Tagen (alle anderen)

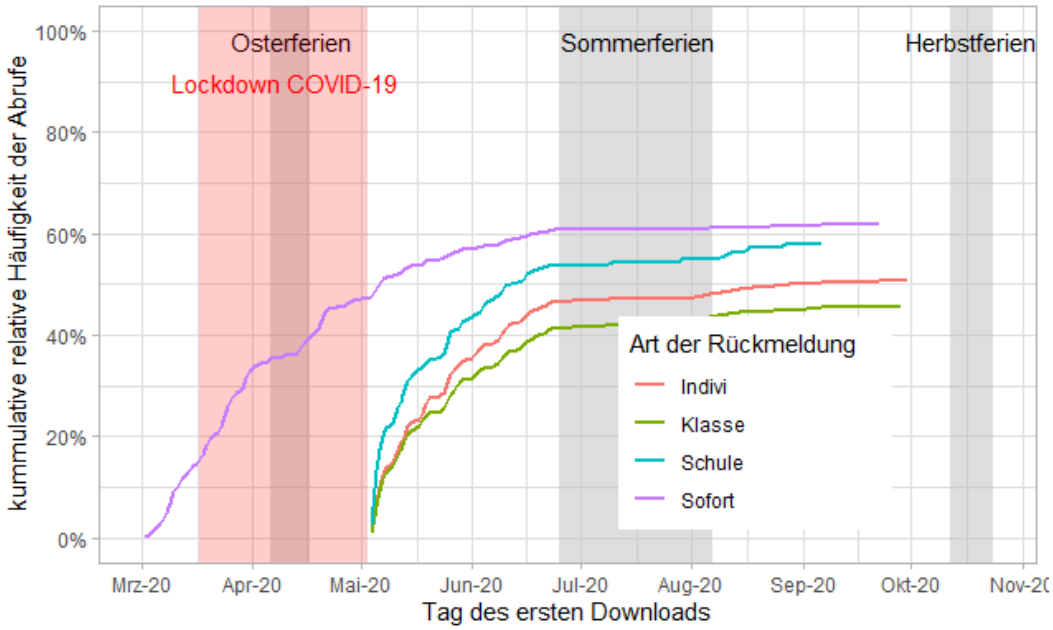


Abbildung 6.5.: Downloads von Rückmeldungen für Mathematik, kumulativ

Tabelle 6.6.: Downloads von Rückmeldungen für das Fach Mathematik, nichtlinear modelliert

Art	nichtlineare Regression Para	est	st.err	t	sig	Anzahl Tage abs	[%]	Ausschöpfung [%]	CI(95%)
Sofort	a	0.6472	0.0039	164.35	0.000	119	50.2	64.4	0.37
	s	-0.0222	0.0003	-63.88	0.000				
Indivi	a	0.4923	0.0027	185.19	0.000	87	50.0	49,2	0.26
	s	-0.0503	0.0010	-48.72	0.000				
Klasse	a	0.4413	0.0031	144.34	0.000	84	48.3	44,1	0.30
	s	-0.0497	0.0013	-38.75	0.000				
Schule	a	0.5475	0.0086	63.42	0.000	119	31.0	54,7	0.86
	s	-0.0682	0.0037	-18.45	0.000				

Tabelle 6.7.: Downloads für das Fach Deutsch

Art	Rückmeldungen verfügbar (Anzahl)	Ende der Eingabe (in %)	von Ausschöpfung 50% erreicht (Datum)	(Tage)	90% erreicht (Datum)	(Tage)	Ausschöpfung (Anzahl)	(in %)
Sofort	1.877	36,5	27.03.	21	28.05.	83	1.139	60,7
Indivi	1.877	-	20.05.	16	22.06.	49	930	49,5
Klasse	1.877	-	19.05.	15	22.06.	49	847	46,6
Schule	445	-	13.05.	9	19.06.	46	283	63,6

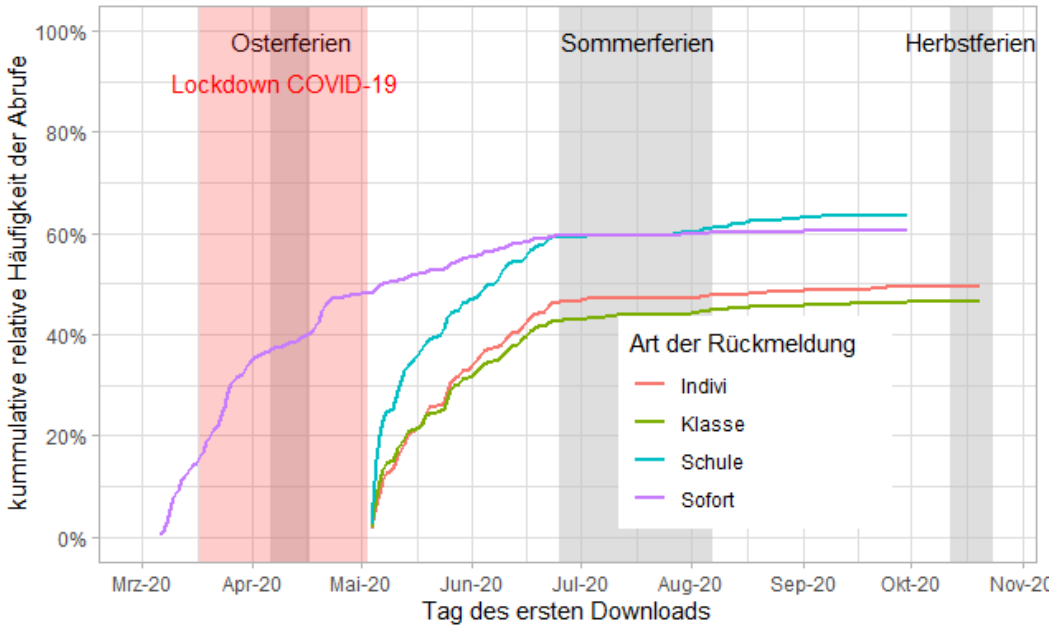


Abbildung 6.6.: Downloads von Rückmeldungen für das Fach Deutsch, kumulativ

Tabelle 6.8.: Downloads von Rückmeldungen für das Fach Deutsch, nichtlinear modelliert

Art		nichtlineare Regression				Anzahl Tage		Ausschöpfung	
	Para	est	st.err	t	sig	abs	[%]	[%]	CI(95%)
Sofort	a	0.6041	0.0027	227.13	0.000	110	47.2	60.4	0.26
	s	-0.0297	0.0004	-74.44	0.000				
Indivi	a	0.4890	0.0029	171.47	0.000	83	47.7	48,9	0.28
	s	-0.0460	0.0009	-49.41	0.000				
Klasse	a	0.4555	0.0038	120.99	0.000	82	47.1	45,5	0.37
	s	-0.0485	0.0014	-33.63	0.000				
Schule	a	0.6009	0.0089	67.57	0.000	60	34.5	60,1	0.89
	s	-0.0671	0.0036	-18.88	0.000				

Tabelle 6.9.: Downloads für das Fach Englisch

Art	Rückmeldungen verfügbar (Anzahl)	Ende der Eingabe (in %)	von Ausschöpfung 50% erreicht (Datum)	(Tage)	90% erreicht (Datum)	(Tage)	Ausschöpfung (Anzahl)	(in %)
Sofort	1.920	41,9	31.03.	27	03.06.	91	1.483	77,2
Indivi	1.920	-	23.05.	19	22.06.	49	1.149	59,8
Klasse	1.920	-	19.05.	15	22.06.	49	1.054	54,9
Schule	439	-	14.05.	10	19.06.	46	268	61,0

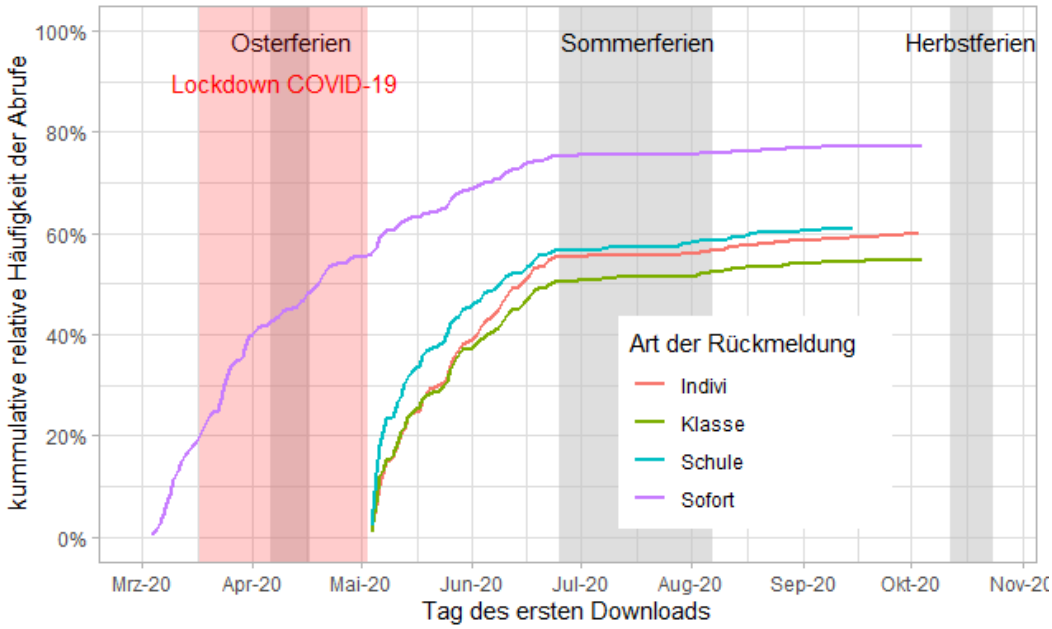


Abbildung 6.7.: Downloads von Rückmeldungen für das Fach Englisch, kumulativ

Tabelle 6.10.: Downloads von Rückmeldungen für das Fach Englisch, nichtlinear modelliert

Art		nichtlineare Regression				Anzahl Tage		Ausschöpfung	
	Para	est	st.err	t	sig	abs	[%]	[%]	CI(95%)
Sofort	a	0.7876	0.0031	253.41	0.000	129	54.9	78.4	0.29
	s	-0.0230	0.0003	-90.83	0.000				
Indivi	a	0.5885	0.0034	171.79	0.000	90	51.7	58,8	0.34
	s	-0.0437	0.0009	-49.59	0.000				
Klasse	a	0.5369	0.0034	156.60	0.000	90	51.7	53,7	0.34
	s	-0.0472	0.0011	-43.75	0.000				
Schule	a	0.5782	0.0078	74.53	0.000	58	33.3	57,8	0.78
	s	-0.0659	0.0031	-21.32	0.000				

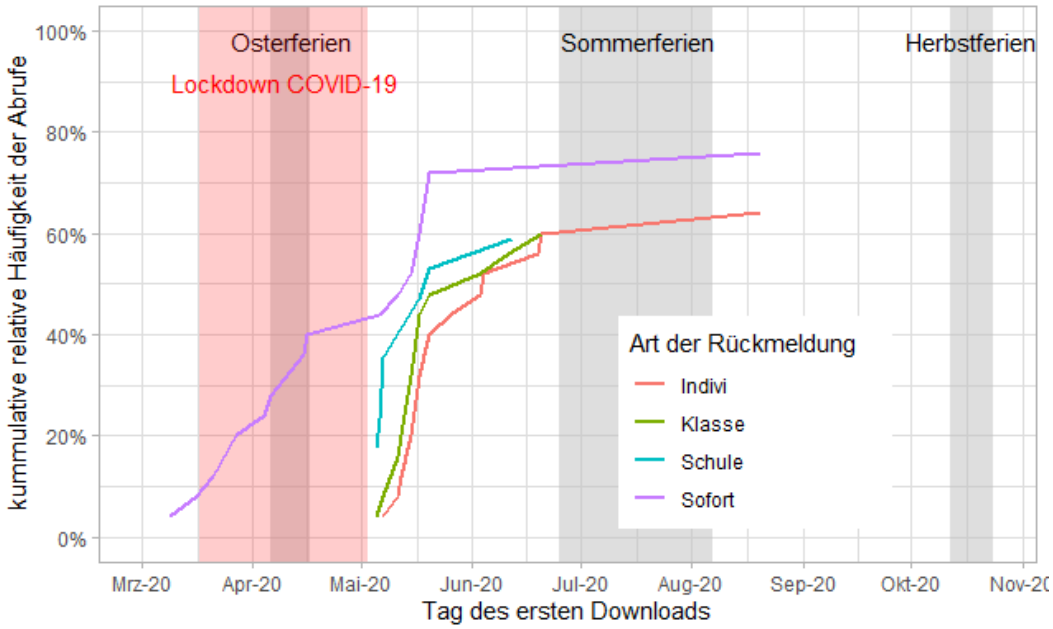


Abbildung 6.8.: Downloads von Rückmeldungen für das Fach Französisch, kumulativ

Tabelle 6.11.: Downloads von Rückmeldungen für das Fach Französisch, nichtlinear modelliert

Art		nichtlineare Regression				Anzahl Tage		Ausschöpfung	
	Para	est	st.err	t	sig	abs	[%]	[%]	CI(95%)
Sofort	a	0.8647	0.0882	9.80	0.000	12	5.2	83.7	6.97
	s	-0.0150	0.0027	-5.61	0.000				
Indivi	a	0.6367	0.0289	22.03	0.000	10	5.8	63,7	2.89
	s	-0.0585	0.0066	-8.80	0.000				
Klasse	a	0.6117	0.0429	14.26	0.000	8	4.6	61,2	4.29
	s	-0.0769	0.0131	-5.87	0.000				
Schule	a	0.5050	0.0503	10.03	0.000	5	2.9	50,5	5.03
	s	-0.5824	0.2536	-2.30	0.070				

ähnlichen Verläufen zu folgen scheinen.

6.5.2. Ergebnisse zu den Forschungsfragen

Zur Untersuchung der Forschungsfragen (vergleiche Abschnitt 6.3) werden die zeitlichen Verläufe der Abrufe für die unterschiedlichen Teilgruppen zuerst graphisch dargestellt. Die beobachteten Merkmale fließen dann in eine logistischen Regression ein, um die Effekte der unabhängigen Einflussgrößen zu analysieren. Für eine bessere Lesbarkeit des Textes sind die Analyseergebnisse im Text für das Fach Mathematik vollständig ausgeführt, werden aber für alle drei Fächer diskutiert. Im Anhang A.7 sind die entsprechenden Ergänzungen für Deutsch und Englisch zu finden. Wegen der geringen Beteiligung an Französisch als erster Fremdsprache, bleibt dieses Fach auch hier unberücksichtigt.

Forschungsfrage 1: Unterscheiden sich die Abrufquoten von Rückmeldungen zwischen Gymnasien und nicht-gymnasialen Schulformen?

Die Verläufe der Abrufe in Abbildung 6.9 lassen sofort erkennen, dass jede Rückmeldung an Gymnasien deutlich häufiger abgerufen wird, als an den nicht-gymnasialen Schulformen. Die Ausschöpfung der Rückmeldeabrufe liegt an den Gymnasien zwischen 66% für die schulbezogene und ca. 95% für die Sofortrückmeldung und damit deutlich höher als an anderen Schulformen. Hier übersteigt die Ausschöpfung für keine Rückmeldung den Wert von 50% deutlich. Die Steigungen sind wie schon zuvor für die einzelnen Rückmeldungen verschieden, aber zwischen den Schulformen etwa gleich.

Auch die logistischen Regressionen (Tabelle 6.12) zeigen, dass die Abrufe sämtlicher Rückmeldungen zwischen gymnasialen und nicht-gymnasialen Schulformen unterschiedlich sind und zwar in jedem Fall so, dass Rückmeldungen an Gymnasien häufiger abgerufen werden[10]. Der Effekt des Einflusses der Schulform kann nach Cohen (1988) für die Sofortrückmeldung als mittel groß und für die anderen Rückmeldungen als klein angegeben werden.

Für die Fächer Deutsch und Englisch stellen sich die Abrufe von Gymnasien und nicht-gymnasialen Schulformen nahezu identisch dar (siehe Graphiken und Ergebnisse der logistischen Regression im Anhang unter A.7.3). Im Fach Deutsch sind selbst die Effektgrößen fast identisch. Nur die Schulrückmeldung übersteigt knapp die 50%-Marke bei den Abrufen an nicht-gymnasialen Schulen, alle anderen Rückmeldungen liegen darunter. An den Gymnasien

[10] Wenn der Koeffizient e^{est} für einen Parameter über 1 liegt und dies auch für das vollständige Konfidenzintervall gilt, dann beeinflusst dieser Parameter die Abrufe dahingehend, dass sie bei größeren Werten häufiger abgerufen werden. Im vorliegenden Fall ist das Gymnasium mit 1 und die anderen Schulformen mit 0 kodiert und alle Konfidenzintervalle für die vier Rückmeldearten liegen vollständig über 1.

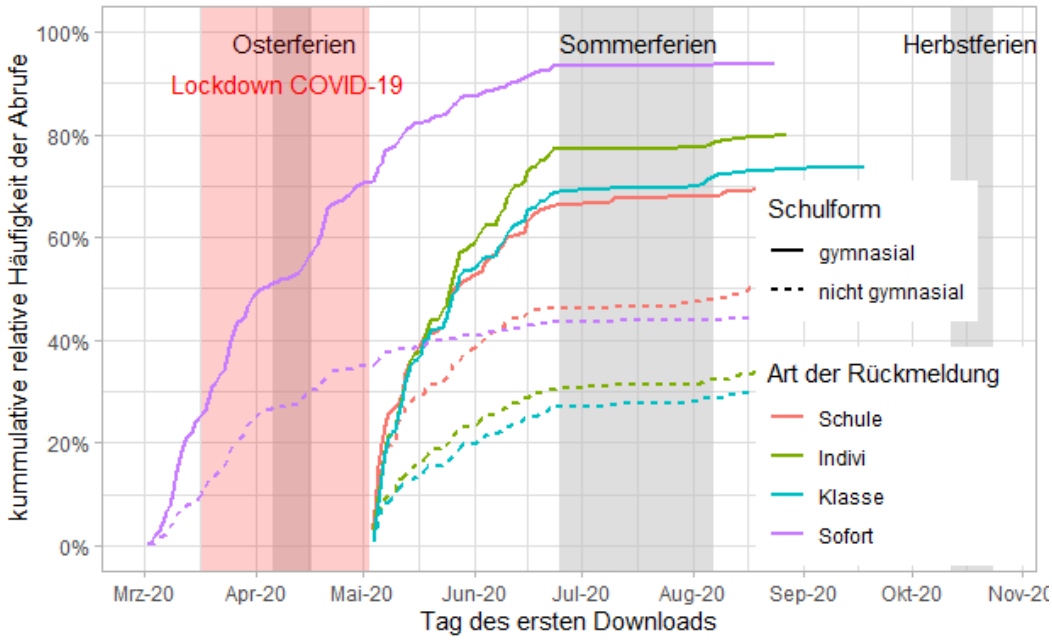


Abbildung 6.9.: Downloads von Rückmeldungen für das Fach Mathematik, nach Schulform

Tabelle 6.12.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Mathematik für verschiedene Schulformen

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-0.2110	0.0559	-3.7737	0.0002	0.8098	[0.73; 0.90]
	Schulform	2.9636	0.1705	17.3814	0.0000	19.3667	[14.04; 27.43]
	R^2	Cohens d			Deviance	dof	AIC
	0,2395	0,5611			2089,82	1976	2093,82
Indivi	(Intercept)	-0,6085	0,0582	-10,4572	0,0000	0,5442	[0,49; 0,61]
	Schulform	1,9930	0,1119	17,8161	0,0000	7,3375	[5,91; 9,16]
	R^2	Cohens d			Deviance	dof	AIC
	0,1736	0,4584			2364,50	1976	2368,50
Klasse	(Intercept)	-0,8115	0,0602	-13,4713	0,0000	0,4442	[0,39; 0,50]
	Schulform	1,8487	0,1058	17,4718	0,0000	6,3513	[5,17; 7,83]
	R^2	Cohens d			Deviance	dof	AIC
	0,1595	0,4355			2383,53	1976	2387,53
Schule	(Intercept)	0.0502	0.1198	0.4190	0.6752	1.0515	[0.83; 1.33]
	Schulform	0.7716	0.2063	3.7394	0.0002	2.1632	[1.45; 3.26]
	R^2	Cohens d			Deviance	dof	AIC
	0,0430	0,2119			592,13	444	596,13

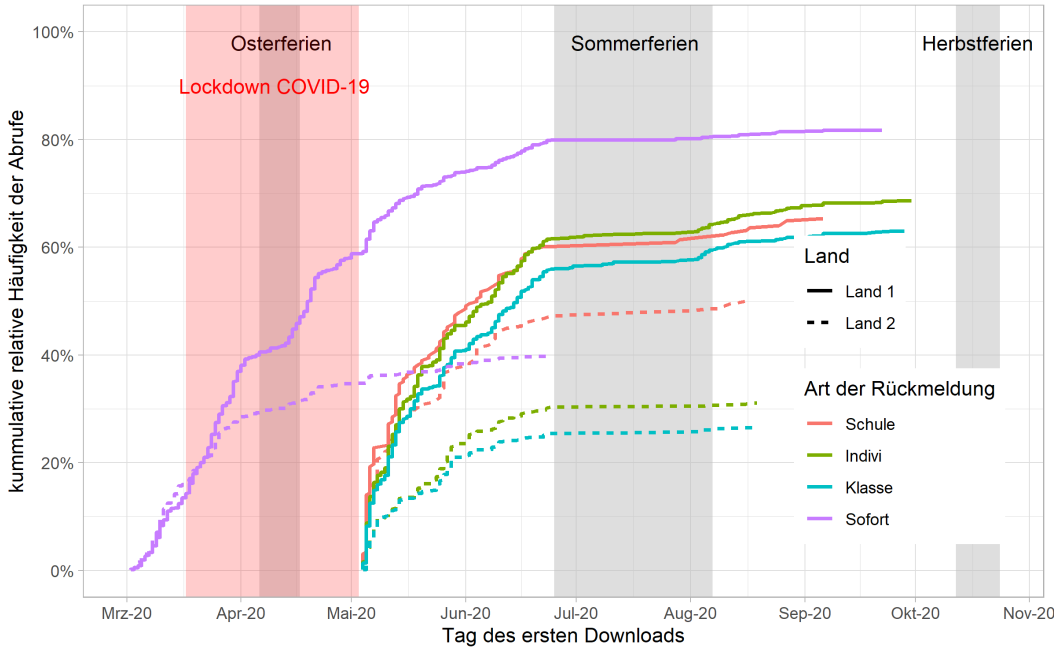


Abbildung 6.10.: Downloads von Rückmeldungen für das Fach Mathematik, nach Land

liegen sämtliche Ausschöpfungen über 70%. Die Schulformunterschiede für das Fach Englisch sind ähnlich, wenn auch nicht von der gleichen Deutlichkeit. So sind alle Effekte nur als klein einzustufen.

Insgesamt lässt sich resümieren, dass der Einfluss der Schulform auf die Abrufe von Rückmeldungen deutlich ist und die Abrufe von Gymnasien dabei häufiger sind, als die anderer Schulformen.

Forschungsfrage 2: Unterscheiden sich die Abrufquoten für Schulen aus den zwei Ländern?

Auch hier finden sich Unterschiede zwischen den zwei Ländern. Die Schulen aus Land 1 rufen die Rückmeldungen für das Fach Mathematik deutlich häufiger ab, als die Schulen aus Land 2. Dramatischer noch ist allerdings der absolute Unterschied beider Länder: Für die drei klassenbezogenen Rückmeldungen liegt die Ausschöpfung Schulen aus Land 1 zwischen 60 und 90 Prozent, in Land 2 hingegen nur bei etwa der Hälfte. Dabei ist eine Abrufquote von 60% der für das Ziel der Unterrichtsentwicklung als zentral eingeschätzte Klassenrückmeldung im Land 1 schon beunruhigend, die 26% für Schulen des Landes 2 hingegen kaum weniger als desaströs. Auch wenn es Möglichkeiten gibt, die Ergebnisse der Vergleichsarbeiten auf anderem Wege an Schüler*innen und Eltern zurückzumelden, als durch die Aushändigung der zur Verfügung gestellten Individualrückmeldung, muss angenommen werden, dass ein

Tabelle 6.13.: Finale Ausschöpfung nach Schulform und Land, in Prozent

Rückmeldung	Land 1		Land 2	
	Gym	nicht-Gym	Gym	nicht-Gym
klassenbezogene (mean)	85,1	62,6	79,2	11,9
Sofort	92,7	75,1	95,8	15,1
Indivi	83,0	59,8	75,8	11,3
Klasse	79,5	52,9	66,1	9,2
Schule	74,7	59,1	63,2	43,7

Drittel bzw. zwei Drittel der Schüler*innen keine solche Rückmeldung erhalten, obwohl deren Ausgabe obligatorisch ist. Die Ergebnisse der logistischen Regression für alle drei Fächer finden sich im Anhang und bestätigen die Interpretation für das Fach Deutsch in gleicher Form. Für das Fach Englisch sind die Unterschiede deutlich kleiner, so dass der Effekt nahezu verschwindet.

Ein Modell unter Einbezug von Schulform und Land

Nach der Erkenntnis, dass einerseits über beide Länder hinweg die Abrufquote an Gymnasien deutlich über jener der anderen Schulformen liegt und zudem über alle Schulformen hinweg in Land 1 deutlich mehr Rückmeldungen abgerufen werden als in Land 2, ergibt sich die Frage, wie sich beide Parameter parallel auf die Abrufhäufigkeit auswirken. Die Tabelle 6.13 zeigt die finalen Ausschöpfungen zum Ende des Untersuchungszeitraumes für die Untergruppen. Mit der Ausnahme der Sofortrückmeldung bei Gymnasien liegen in jedem Vergleich die Ausschöpfungen der Schulen aus Land 1 über denen aus Land 2 und die der Gymnasien über denen der nicht-Gymnasien.

Diese Gegenüberstellung wird durch eine logistische Regression ergänzt, die Land und Schulform gleichermaßen einbezieht, so dass auch das Verhältnis beider Effekte abgeschätzt werden kann. Die umfassenden Ergebnisse der einzelnen Regressionen für jedes Fach und jede Rückmeldung sind den Tabellen im Anhang zu entnehmen (Abschnitt A.7.5). Fazit: Beide Parameter tragen signifikant und unabhängig voneinander zur Erklärung der Rückmeldeabrufe bei.

Forschungsfrage 3: Gibt es einen Zusammenhang zwischen der durchschnittlich erreichten Leistung und den Abrufquoten?

Für die explorative Untersuchung dieser Forschungsfrage, werden die Fachleistungen der Schulen am Mittelwert in zwei Gruppen geteilt und untersucht, in wieweit die Zugehörigkeit zu den Gruppen den Abruf der Rückmeldung beeinflusst[11]. Die Ergebnisse der logistischen Regression sind folgend für das Beispiel der schulbezogenen Rückmeldung des Faches Mathematik in Tabelle 6.14 (erste Regression, oben) wiedergegeben. Es findet sich ein sehr kleiner Effekt. Es muss aber vermutet werden, dass dieser alternativ durch die Schulform erklärt werden kann. Denn zuvor wurde festgestellt, dass die Gymnasien mit den durchschnittlich besseren Ergebnissen die Rückmeldungen häufiger abgerufen haben. Deshalb wurde das Modell zuerst um die Schulform (zweite Regression) und abschließend noch um das Land (dritte Regression) erweitert. Auch diese Ergebnisse sind der Tabelle zu entnehmen.

Durch den Einbezug der Schulform wird der Effekt der Leistungsunterschiede vollständig subsumiert. Als letztes wurde das Modell aus dem vorhergehenden Abschnitt wiederholt, welches nur Schulform und Land einbezieht und den Parameter der Leistung vernachlässigt. Dieses letzte Modell der Tabelle 6.14 ist entsprechend des AIC allen anderen vorzuziehen. Dies trifft bis auf einen einzigen Fall[12] auch auf die Modellierungen aller Fächer und Rückmeldungen zu. Im Abschnitt A.7.6 des Anhangs sind die Ergebnisse dieser Berechnungen vollständig wiedergegeben. Die Leistung, hier sehr einfach mit einer Gruppierung in zwei Leistungsgruppen am Mittelwert modelliert wurde, trägt nicht signifikant zur Aufklärung der Abrufe bei.

6.6. Diskussion

6.6.1. Befunde der deskriptiven Analyse

Der Untersuchungszeitraum bis zum Ende der Herbstferien (25. Oktober 2020) ist offensichtlich ausreichend gewählt: Die Kurven brechen bei der letzten erfolgten Rückmeldung innerhalb dieses Zeitraumes ab. Für die Sofortrückmeldung des Fachs Mathematik in Brandenburg erkennt man in der Abbildung 6.10 auf Seite 194, dass der letzte Abruf kurz vor

[11] Um die logistischen Regressionen mit den zuvor gerechneten Modellen vergleichen zu können, muss hier auf eine Leistungsvariable zurückgegriffen werden, die für beide Länder vollständig vorliegt. Da die Teilnahme an Deutsch Lesen und Englisch Hörverstehen in Brandenburg freiwillig war, wurden als Fachleistungen die Orthographieergebnisse für Deutsch und die Ergebnisse des Leseverstehens für Englisch genutzt.

[12] Nur bei der Individualrückmeldung für das Fach Englisch ist dem Modell unter Einbezug der Leistungsvariable auf der Basis des AIC-Vergleichs (2.483 zu 2.485) knapp der Vorzug zu geben.

Tabelle 6.14.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Mathematik für zwei verschiedene Leistungsgruppen

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Schule	(Intercept)	0.1071	0.1285	0.8336	0.4045	1.1130	[0.8655; 1.4331]
	Leistung	0.4914	0.1950	2.5200	0.0117	1.6347	[1.1173; 2.4013]
	R^2	Cohens d			Deviance	dof	AIC
	0,0192	0,1400			600,19	444	604,19
Schule	(Intercept)	0.0845	0.1289	0.6625	0.5077	1.0892	[0.8461; 1.4035]
	Leistung	-0.2341	0.3164	-0.7398	0.4594	0.7913	[0.4217; 1.4662]
	Schulform	0.9624	0.3309	2.9086	0.0036	2.6179	[1.3759; 5.0613]
	R^2	Cohens d			Deviance	dof	AIC
	0,0446	0,2160			591,58	443	597,58
Schule	(Intercept)	-0.2114	0.1635	-1.2926	0.1961	0.8095	[0.59; 1.11]
	Leistung	-0.1823	0.3217	-0.5668	0.5708	0.8333	[0.4397; 1.5612]
	Schulform	0.9045	0.3363	2.6894	0.0072	2.4708	[1.2844; 4.8259]
	Land	0.5910	0.1975	2.9920	0.0028	1.8058	[1.2279; 2.6652]
	R^2	Cohens d			Deviance	dof	AIC
	0,0707	0,2758			582,54	442	590,54
Schule	(Intercept)	-0.2419	0.1545	-1.5652	0.1175	0.7851	[0.5787; 1.0617]
	Schulform	0.7553	0.2084	3.6239	0.0003	2.1282	[1.4201; 3.2182]
	Land	0.5973	0.1972	3.0289	0.0025	1.8172	[1.2365; 2.6804]
	R^2	Cohens d			Deviance	dof	AIC
	0.0698	0.2739			582.86	443	588.86

den Sommerferien erfolgte. Für die Mehrzahl der Rückmeldungen erfolgt der letzte Abruf im Untersuchungszeitraum bis Anfang Oktober. Ca. einen Monat nach dem Untersuchungszeitraum öffnete dann schon das Portal des folgenden Durchgangs von VERA-8. Dass im Untersuchungszeitraum einige Abrufe für zurückliegende Projekte verzeichnet wurden lässt vermuten, dass auch zu deutlich späteren Zeitpunkten weitere Downloads von Rückmeldungen erfolgen. Solche späten Abrufe können keine Relevanz für Unterrichtsentwicklung im Sinne eines formativen Assessments entfalten und bleiben hier unbeachtet. Deren Zahl scheint aber auch vernachlässigbar klein.

In den Betrachtungen der Abrufe über den Zeitverlauf sind die Sofortrückmeldungen von den späteren Rückmeldungen zu unterscheiden. Von jenen Lehrkräften, die letztere Rückmeldungen tatsächlich abrufen, nutzt ein großer Teil den Zeitraum direkt nach der Freischaltung. Für die verschiedenen Rückmeldungen der unterschiedlichen Fächer erfolgen 50% aller Abrufe in den ersten ein bis drei Wochen (9 bis maximal 19 Tage) und mehr als 90% bis zu den Sommerferien (52 Tage). Nimmt man die Präsenzwoche (letzte Woche der Sommerferien)

aus, werden während der 36 Tage Sommerferien nicht mehr als 3% der Rückmeldungen das erste Mal heruntergeladen. Aber auch im neuen Schuljahr werden inkl. der Präsenzwoche bis zum Ende der Herbstferien (72 Tage) maximal 4% abgerufen.

Wichtig ist, dass hier wie im Weiteren immer nur die jeweils ersten Abrufe von Rückmeldungen analysiert werden. Eine Schlussfolgerung, dass nach dem Ende des Schuljahres die Rückmeldungen nicht weiter vorgehalten werden müssten, wäre demnach falsch. Tatsächlich wird jede Rückmeldung im Schnitt 2,3 Mal heruntergeladen. Es ist also durchaus möglich, dass Lehrkräfte, die eine 9. Klasse neu übernehmen, die Rückmeldungen erneut aufrufen, beispielsweise um sich mit den Ergebnissen einen Überblick über den Leistungsstand der Klasse insgesamt oder der einzelnen Schülerinnen und Schüler zu verschaffen. Das kann allerdings nicht über die Downloads analysiert werden, denn eine Weitergabe der Rückmeldung ist ebenso denkbar, vermutlich sogar wahrscheinlicher.

Die Sofortrückmeldungen nehmen eine besondere Rolle ein. Sie werden nur wenige Minuten nach der Eingabe der Ergebnisse vom ISQ-Portal automatisch zur Verfügung gestellt, und damit zu einem Zeitpunkt, zu dem die Lehrkräfte das Portal ohnehin geöffnet haben. Es kann angenommen werden, dass die Ausschöpfung deshalb höher ist, als die der anderen drei erst zu einem späteren Zeitpunkt freigeschalteten Rückmeldungen. Allerdings können die Rückmeldungen auch nicht unmittelbar nach der Eingabe der Ergebnisse angesehen werden, denn bis zur Freischaltung der Sofort-Rückmeldung vergehen nach dem Abschluss der Eingaben für eine Lerngruppe aus technischen Gründen zwischen 3 und 8 Minuten. Die Ausschöpfung bis zum Ende der Eingabezeit, wenn also fast alle Ergebnisse eingegeben sind, liegt dann doch nur zwischen 35 und 42%. Deutlich mehr als die Hälfte der Lehrkräfte warten nach abgeschlossener Eingabe der Ergebnisse diese wenigen Minuten also nicht ab, um die Sofort-Rückmeldung gleich abzurufen. Dies könnte gleichermaßen ein Beleg für Unkenntnis oder Desinteresse sein.

Grundsätzlich werden selbst individuelle Rückmeldungen nur zwischen 50 und 60% bis zu den Herbstferien abgerufen. Natürlich ist es grundsätzlich möglich, aus den Klassenrückmeldungen oder den Rohdaten individuelle Ergebnisse abzuleiten und mit diesen der als obligatorisch festgelegten Rückmeldung an Schüler*innen und Eltern zu genügen. Allerdings werden auch diese zu beträchtlichen Teilen nicht abgerufen. Ebenso kann nicht ausgeschlossen werden, dass die Lehrkräfte aus den Korrekturen der Testhefte Ergebnisse ableiten und den Schüler*innen zur Verfügung stellen. Erwartet werden könnte dies zum Beispiel an Schulen, in denen die Vergleichsarbeiten regelwidrig und ggf. ohne Rücksicht auf die getesteten Inhalte,

wie eine Klassenarbeit benotet werden. Im ungünstigsten Fall muss aber davon ausgegangen werden, dass ein bis zwei Drittel aller Schülerinnen und Schüler keine bzw. keine im Portal erstellte individuelle Rückmeldung über ihre Ergebnisse erreicht.

Die klassenbezogenen Rückmeldungen sind aus der Sicht des ISQ die zentralen Ergebnisrückmeldungen, mit denen Lehrkräfte weiterarbeiten sollen. Die mitgelieferten Materialien, zur Verfügung gestellte digitale Tools wie auch Workshops sollen dies unterstützen. Im Gegensatz zur obligatorischen Ausgabe einer individuellen Rückmeldung besteht hier aber keine explizit formulierte Pflicht, die einen Abruf im Allgemeinen erforderlich macht. So liegen die Abrufquoten dieser Rückmeldungen auch unter denen der individuellen Rückmeldungen, allerdings nur geringfügig. Die Nutzung des Aufgabenbrowsers (ISQ, 2021) als einem unterstützenden Tool weist zu zwei Zeitpunkten eine größere Nutzungshäufigkeiten auf. Direkt vor den Vergleichsarbeiten werden offenbar Aufgaben oder Aufgabenhefte heruntergeladen, um mit den Schülerinnen und Schüler vorbereitend zu üben bzw. sich ein Bild von den zu erwartenden Anforderungen an die Schüler*innen zu machen. Aber auch nach den Vergleichsarbeiten findet sich stets eine verstärkte Nutzung. Über die direkte Verlinkung von Aufgabenstatistiken aus den Rückmeldungen ergeben sich beispielsweise unterschiedliche Anlässe, die eine gezielte Weiterarbeit vermuten lassen. Hier bietet sich der Anschluss weitergehender Untersuchungen an.

Schulbezogene Rückmeldungen werden etwas häufiger, vor allem aber schneller abgerufen. Die vergleichende Übersicht der Ergebnisse der unterschiedlichen Lerngruppen und der Bezug zu einem sozial adjustierten Vergleich ist für Fachkoordinator*innen wie für die Schulleitungen sicher von Interesse. Eine Ausschöpfung von 58 bis 64% kann trotzdem nicht zufriedenstellen.

Fazit: Allein die Sofortrückmeldung der Fremdsprachen bildet mit 77,2% eine Ausnahme, die Ausschöpfungen liegen sonst länder- und schulformübergreifend nur zwischen 45,7% und 63,6%. Dies macht den Grad an Antworten im Sinne sozialer Erwünschtheit deutlich, die in Studien anzeigen, dass beträchtliche Teile der Lehrerschaft auf der Basis der VERA-Rückmeldungen zu einer Weiterarbeit animiert worden sein wollen. Zumindest für die Länder Berlin und Brandenburg, aber auch für Thüringen muss dies in Zweifel gezogen werden. Allerdings gibt es wenig Gründe dafür anzunehmen, dass die Quoten für den Abruf der Rückmeldungen und damit der minimalen Anforderung für eine erfolgreiche Weiterarbeit in anderen Ländern deutlich besser ausfallen. Das ISQ, als das für die administrative Durchführung der Vergleichsarbeiten verantwortliche Institut, unternimmt wie andere Länder einige Anstrengungen, um die Idee hinter den Vergleichsarbeiten zu transportieren. Auch die gerin-

ge Einbindung von VERA in andere Kontexte schulischer Qualitätssicherung ist sicher kein Alleinstellungsmerkmal Berlins und Brandenburgs.

Die Betrachtung der Abrufe über die Zeit mit Hilfe der nichtlinearen Regression dokumentiert die gerade referierten Effekte. Der Parameter der Ausschöpfung beschreibt in der Modellfunktion jenen Wert, dem sich die Abrufquote zeitlich asymptotisch nähert. Ein leichter Anstieg nach den Sommerferien lässt die reale Abrufquote diesen Wert in einigen Fällen geringfügig überschreiten, umgekehrtes gilt für die Abflachung der Abrufe mit dem Start der Sommerferien. Für den Untersuchungszeitraum ist dieser Wert aber ebenso eine angemessene Näherung für die Ausschöpfung, wie der durch die Regression für das Ende des Untersuchungszeitraumes geschätzte Wert. Die Regressionen zeigen, dass die Modellfunktion die realen Verläufe sehr gut abbilden. Lediglich für Französisch sind die Werte problematisch. Hier erfolgen aber auch nur an sehr wenigen Tagen Abrufe (siehe Tabelle 6.11 *Anzahl Tage abs*). Die sehr gute Annäherung durch die Modellfunktion zeigt sich auch im kleinen Konfidenzintervall für die Schätzung der Ausschöpfung zum Ende des Untersuchungszeitraumes von im Allgemeinen weniger als einem Prozent.

Für eine Interpretation des Parameters für die Steigung kann man sich folgender Hilfe bedienen: Im Fall der Sofortrückmeldung für das Fach Mathematik ist die Steigung gleich $s = -0.0222$. Am Tag $1/s$, also hier am Tag 44, ergibt sich als Potenz -1. Die kumulative Ausschöpfung beträgt an diesem Tag genau $1 - e^{-1} = 63,2\%$. Allgemein gilt: Der Tag, der sich aus dem Reziprok der Steigung ergibt beschreibt jenen Punkt im Funktionsverlauf, an dem 63,2% der asymptotischen Ausschöpfung erreicht werden. Die Tabelle 6.15 gibt einen Überblick über die sich hieraus ergebenden Werte, die nun einen alternativen Vergleich der Steigungen, also der Geschwindigkeit der Abrufe zulassen. Die Werte für das Fach Französisch sind hier nur aus Gründen der Vollständigkeit ergänzend aufgeführt, weil die geringe Zahl der Messpunkte keine sichere Modellierung erlaubt. Dass die Sofortrückmeldung die mit Abstand geringste Steigung aufweist wundert nicht. Bei dieser Rückmeldung liegen nicht alle Rückmeldungen am ersten Tag möglicher Abrufe vor, sondern immer nur die zuvor Eingegebenen. Die größere Steigung im Fach Deutsch könnte allerdings darauf hinweisen, dass die Deutsch-Lehrkräfte die Ergebnisse schneller eingeben und damit auch die Rückmeldungen schneller abrufen können, oder aber, dass sie tatsächlich die Rückmeldungen schneller abrufen. Alle folgenden Rückmeldungen stehen zu einem konkreten Zeitpunkt zur Verfügung. Man kann natürlich davon ausgehen, dass eine Lehrkraft, die sich in das ISQ-Portal einwählt, um Rückmeldungen anzusehen, alle verfügbaren Rückmeldungen abruft. Damit würden sich

Tabelle 6.15.: Anzahl der Tage bis 63,2 Prozent der asymptotischen Ausschöpfung erreicht sind

Rückmeldung	Mathematik	Deutsch	Englisch	Mittelwert[a]	Französisch
Sofort	45.0	33.7	43.5	40.7	66.7
Indivi	19.9	21.7	22.9	21.5	17.1
Klasse	20.1	20.6	21.2	20.6	13.0
Schule	14.7	14.9	18.2	15.9	1.7

[a]Wegen der kleinen Zahlen und der damit verbundenen Ungenauigkeit wurden die Werte für Französisch im Mittelwert nicht berücksichtigt.

nicht nur bezüglich der Abrufzahlen sondern auch der Abrufgeschwindigkeit sehr ähnliche Werte ergeben. Für die klassenbezogene Rückmeldung und jene für die Weitergabe an die Schülerinnen und Schüler trifft dies zu. Die Zugriffe auf die schulbezogenen Rückmeldungen erfolgen allerdings zügiger. Diese Zahlen bestätigen damit, was sich aus den Graphiken ablesen lässt.

6.6.2. Befunde zu den Forschungsfragen

Um die Ergebnisse der Abrufquoten zu plausibilisieren, fanden sich bis auf die Thüringer Ergebnisse keine anderen Veröffentlichungen mit Analysen von Rückmeldungen aus Vergleichsarbeiten, beispielsweise zur differenten Rezeption für verschiedene Schulformen oder in unterschiedlichen Ländern. Allerdings wird für das Land Brandenburg im Rahmen der die Schulvisitation vorbereitenden Befragungen von Lehrkräften eine Frage zum Umgang mit den Vergleichsarbeiten gestellt. In ihrem Bericht über die zweite Runde der Schulvisitationen aller Schulen berichten Eiben und Preuße (2017) die Ergebnisse der einzelnen Fragen. Zur Thematik des Umgangs mit den Ergebnissen von Vergleichsarbeiten finden sich dort drei Fragen, von denen sich aber nur Frage 43 ausschließlich auf die Vergleichsarbeiten bezieht[13]:

> 43. In der Konferenz der Lehrkräfte werten wir regelmäßig die Ergebnisse der Vergleichsarbeiten (Jahrgangsstufen 3, 6 oder 8) aus.

Die Frage bezieht sich nur auf eine besondere Form der Auswertung. Trotzdem wäre zu vermuten, dass sich hierbei die Unterschiede zwischen den Schulformen im Land Brandenburg widerspiegeln. Die Lehrkräfte konnten hier *überwiegend schwach* (1), *eher schwach als stark*

[13]Die Schulen Brandenburgs haben zum Zeitpunkt der zweiten Runde der Schulvisitation neben den bundesweit verbindlichen Vergleichsarbeiten in den Klassenstufe 3 und 8 zusätzlich an den in einigen Ländern eingeführten Vergleichsarbeiten der Klassenstufe 6 teilgenommen. Diese Vergleichsarbeiten sind hier mit aufgeführt. Die anderen zwei Fragen beziehen weitere Aspekte der schulischen Arbeit ein. Welcher Anteil auf die Rezeption der Rückmeldungen von Vergleichsarbeiten fällt, lässt sich nicht feststellen.

(2), *eher stark als schwach* (3) oder *überwiegend stark* (4) zustimmen oder die Frage unbeantwortet lassen. Unter Ausschluss der 13% Unbeantworteten wird im Bericht ein Mittelwert von 3,4 ausgewiesen bzw. eine Zustimmung für die oberen beiden Kategorien von 89%.

Differenziert ausgewertet wurden die Antworten von öffentlichen Schulen mit Sekundarstufe I und hier die Gymnasien und die nicht-gymnasialen Schulformen. Von den knapp je 2.000 Antworten aus jeder Schulform fielen in die oberen zwei Antwortkategorien 85% bei den Gymnasien und 86% bei den anderen. Die Schulform hat hier keinen nachweisbaren Einfluss auf die Auseinandersetzungen mit den Ergebnissen der Vergleichsarbeiten. Allerdings könnte es auch sein, dass eben gerade auf die Beschäftigung in der Konferenz der Lehrkräfte in Brandenburg von schulaufsichtlicher Seite besonderer Wert gelegt wird, was vielleicht zu der konkreten Nachfrage geführt hat. Es könnte aber auch angenommen werden, dass diese Antworten vielleicht insbesondere für die nicht gymnasialen Schulformen im Sinne sozialer Erwünschtheit verzerrt sind.

Mit den Daten der vorliegenden Vollerhebung gilt für alle Rückmeldungen, dass die Chance für einen Abruf an Gymnasien signifikant höher, teilweise deutlich höher ist als an den anderen Schulformen. Die Gründe für die starke Vernachlässigung der Rückmeldungen an nicht-gymnasialen Schulformen müssen dringendst untersucht werden. Das Gesamtbild defizitärer Abrufe geht offensichtlich zu einem beträchtlichen Anteil auf die Vernachlässigung an nicht-gymnasialen Schulformen zurück. Ebenso sind die Abrufzahlen in Land 1 signifikant höher als in Land 2. Die Fachleistung als Einflussgröße wird von der Schulform moderiert. Allerdings liegt den vorliegenden Analysen eine nur sehr einfache Modellierung der Leistung als Einflussgröße zugrunde.

Die über alle Fächer, Domänen und Rückmeldearten unerwartet niedrigen Abrufquoten decken sich grob mit den Berichten aus Thüringen. In Thüringen findet sich kein nennenswert bedeutsamer Einbruch der Abrufzahlen im Zuge der Corona-Pandemie. So kann erwartet werden, dass sich auch die hier gefundenen niedrigen Abrufquoten nicht auf einen Effekt der Corona-Pandemie zurückführen lassen. Ebenso kann vermutet werden, dass die Abrufe auch in Berlin und Brandenburg seit der Einführung gesunken sind und sich nach über 10 Jahren äquivalent zu Thüringen auf dem hier aufgezeigt niedrigem Niveau befinden. Eine solche Entwicklung deckt sich mit den schon zitierten Befunden von Groß Ophoff (2013).

Zudem ergeben sich weitere Ansätze für Forschungsfragen. So wurde im Durchgang 2020 das erste Mal allen Berliner Schulen das Angebot unterbreitet, die sonst als Papiertest ausgeführten Vergleichsarbeiten als Online-Test durchzuführen. Dieses Angebot bezog sich nur

auf die sprachlichen Fächer, also nicht auf Mathematik, und wurde für ca. die Hälfte der Schülerinnen und Schüler umgesetzt. Die Tests fanden in diesem Fall nicht an einem Tag statt, sondern in einem etwa zweiwöchigen Korridor, welcher mit dem ersten Testtag begann. Für die Lehrkräfte erhöht sich bei einem Online-Test der vorgängige Aufwand, wegen der notwendigen Raumplanung und der Vorbereitung der schülerweisen Einwahlcodes. Allerdings erübrigt sich vollständig die Eingabe und zu einem erheblichen Anteil die Korrektur der Antworten der Schüler*innen. Nur einige offene Aufgaben müssen durch die Lehrkräfte in einem diesen Prozess unterstützendem Portal bewertet werden. Dieses Angebot ist die Ausgestaltung eines Beschlusses der Senatsverwaltung für Bildung, Jugend und Familie (2019) mit dem Ziel, Lehrkräfte zu entlasten.

> „Lehrkräfte können sich dadurch stärker auf die Auswertung der Ergebnisse und die Folgen für die Schul- und Unterrichtsentwicklung sowie die Förderung der Schülerinnen und Schüler konzentrieren."

Lehrkräften an Berliner Schulen sollten durch die Teilnahme am Onlinetest mehr Zeit für die Arbeit mit den Rückmeldungen zur Verfügung stehen. Interessant ist hier zu untersuchen, ob die Entscheidung für den Online-Test tatsächlich mit einer stärkeren Beschäftigung mit den Rückmeldungen einherging. Zu vermuten wären höhere Abrufquoten bei Lerngruppen mit Online-Tests. Wegen der weiterhin den Schulalltag bestimmenden Corona-Pandemie wurde die Testteilnahme an VERA-8 2021 freiwillig gestellt. Auch diese veränderten Rahmenbedingungen sollten den Anteil abgerufener Rückmeldungen deutlich erhöhen.

Mit dem Jahr 2021 konnten Berliner Schulen erstmals auf Ebene der Lerngruppe entscheiden, ob mit dem einfacheren, primär für nicht-gymnasiale Schulen entwickeltem Testheft getestet werden soll oder mit dem schwierigeren Testheft, dass eher die Anforderungen im Gymnasium widerspiegeln soll. Interessant könnte die Analyse sein, ob Lerngruppen, die sich nicht für das für ihre Schulform empfohlene Testheft entschieden haben, eher die Rückmeldungen abrufen. Es kann vermutet werden, dass sich in der aktiven Auswahl eines Testhefts ein Engagement ausdrückt, welches sich auch in einer höheren Ausschöpfung der Rückmeldeabrufe niederschlägt. Effekte könnten dabei für Rückmeldungen verschiedener Ebenen unterschiedlich ausfallen.

7. Gesamtdiskussion

Für den Beitrag, den diese Arbeit zur Diskussion von Validität der Vergleichsarbeiten bezüglich ihres Kernziels der Unterrichtsentwicklung leistet, ist die Definition von Validität aus dem Kapitel 2 grundlegend.

> „Validität ist ein integriertes, bewertendes Urteil über das Ausmaß, in dem die Angemessenheit und Güte von Interpretationen und Maßnahmen auf der Basis von Testwerten [...] durch empirische Belege und theoretische Argumente gestützt sind.“(Messick, 1989a, S.13; in der Übersetzung von Hartig et al., 2012, S.144)

Es geht nicht darum, Validität als einen dichotomen Status zuzuweisen, sondern ein „Urteil über [ein] *Ausmaß*“ zu treffen. Und dabei sind es auch die einzelnen Verfahrensschritte der Messung, aber letztendlich essentiell die abgeleiteten Maßnahmen, die sich als strukturell korrekt erweisen müssen. Der Intention dieser Definition folgend sollen alle Betrachtungen der Validität von Vergleichsarbeiten den Anschluss an die intendierte Nutzung im Schulpraktischen suchen. Diese Arbeit untersucht dazu

1. die Bewertung der Testbearbeitung mit Hilfe der Rasch-Skalierung und damit die für die Vergleichsarbeiten konstitutiven psychometrischen Grundlagen,

2. die Gültigkeit der Generalisierung, die sich über die Konsistenz der Metrik der eingesetzten Instrumente manifestiert sowie

3. die Abrufquoten als empirischer Beleg für eine fundamentale Grundlage der Zielerreichung.

In den Standards der American Educational Research Association (AERA) et al. (2014) wird darauf verwiesen, dass im Prozess des Nachweises von Validität die Testentwickler*innen Verantwortung dafür tragen, dass die ihrer Einschätzung nach intendierten Interpretationen zu den beabsichtigten Konsequenzen führen können. Gleich anschließend wird aber auch ausgeführt, dass die Testanwender*innen letztlich die Verantwortung für die Bewertung der

Qualität der vorgelegten Validitätsnachweise und ihrer Relevanz für die lokale Situation tragen (ebenda S.23). Es reicht also nicht, dass Tests psychometrischen Anforderungen genügen und dass ihnen eine schlüssig definierte Fachdidaktik zu Grunde liegt. Es bedarf auch einer schulpraktischen Anschlussfähigkeit. Der oft vorgebrachte Verweis darauf, dass dies die professionelle Arbeit der Lehrkräfte sei und es dazu nur einer angemessenen Erläuterung der Konzepte und ggf. mehr Angebote zur Weiterbildung benötigte, reicht nicht aus. Insofern ist die Implementation der Vergleichsarbeiten bisher unvollständig geblieben.

Diedrich (WZB, 2021, 29:09) korrigiert dazu selbstkritisch ihre frühere Sicht: „Wir haben lange so argumentiert [...]: ‚Um Gottes willen, wir können doch den Schulen nicht erklären, wie sie ihre Daten verwenden. Das wissen die doch viel besser.' Heute halte ich das für falsch. [...] Die Schulen mit Daten allein zu lassen, halte ich für wirklich nicht verantwortlich." Diese Positionsverschiebung ist aus Sicht des Autors geboten. Validität herzustellen bedeutet offensichtlich, hier Verantwortung wahrzunehmen und den Prozess der Qualitätsentwicklung mit Vergleichsarbeiten als Ganzen zu verstehen und als Testentwickler*in und Testadministration zusammen mit der Unterrichtspraxis zu gestalten.

Das beginnt bei einer klaren Zieldefinition. Unbestritten ist dies primär die Nutzung der Ergebnisse zum Zwecke von Unterrichts- und Schulentwicklung. Dass hier weitere Nutzungsräume offengehalten werden, macht eine konkrete Ausgestaltung der Vergleichsarbeiten deutlich schwieriger, als es beispielsweise für den diesbezüglich klar eingeordneten Bildungstrend der Fall ist. Qualitätssicherung und -entwicklung durch Unterrichtsentwicklung auf der Basis von VERA-Rückmeldungen bedeutet, der Nutzung durch die Lehrkräfte in den Schulen unbedingten Vorrang einzuräumen.

7.1. Zusammenfassung zentraler Befunde

Die Befunde dieser Arbeit erlauben gerade keine vollständige Bewertung von Validität. Sie konzentriert sich auf Aspekte, die in der Auseinandersetzung mit dem Gegenstand augenfällig problematisch sind. Ohne diese deutliche Einschränkung ergäbe sich eine unangemessen kritische Einordnung der Vergleichsarbeiten. Bei aller Kritik in dieser Arbeit erweist sich die Implementation der Vergleichsarbeiten immer öfter als überaus hilfreich, ja gelungen. Und dass sich diese Beispiele gerade eben in der Schule, im konkreten Unterricht finden, bestätigt dann auch die Schwerpunktsetzung dieser Arbeit. In den folgenden Abschnitten werden die drei Validitätsaspekte *Skalierung*, *Generalisierung* und *Rückmeldungen* (vergleiche das Kane'sche Validitätsargument, Abbildung 2.1 auf Seite 36) zusammengefasst.

7.1.1. Skalierung

Einige Aussagen zur psychometrischen Testkonstruktion aus dem Kapitel 4 sind grundlegend und wurden bisher als gültig vorausgesetzt. Sie wurden in dieser Arbeit deshalb aus der Sicht der Psychometrie als Gewissheiten bezeichnet und trotzdem aus schulpraktischer Sicht in Frage gestellt. Dabei wurden vier Problemstellen identifiziert.

Die Prozesse der Entwicklung, Pilotierung, Auswahl und Normierung von Items sind für die Vergleichsarbeiten gegenüber den Prozessen beim Bildungstrend kongruent angelegt, lediglich ein stückweit vereinfacht worden. Die plausiblen Ergebnisse beim Bildungstrend und anderen internationalen Studien sprechen für dieses Vorgehen, auch wenn dem Rasch-Modell eine gewisse Einfachheit zugeschrieben werden muss. Es konnte gezeigt werden, dass eine Auswahl von Items mit stärkeren Abweichungen des Infits[1] unter bestimmten Bedingungen zu Verzerrungen der Fähigkeitsmessung führen können. Bei der Itemauswahl wurden Items mit Infit-Werte auch unter 0,9 explizit zugelassene. Diese höhere Trennschärfe wird als positive Modellabweichung akzeptiert. Dies führt allerdings, wie gezeigt wird, bei einem unangemessen leichtem (schwerem) Testheft zur Unterschätzung (Überschätzung) der Fähigkeit. Für den praktisch selteneren Fall, dass der Infit größer als 1,1 wird, gelten inverse Beziehungen. Um solchen Verzerrungen entgegenzuwirken muss eine Itemauswahl deutlich restriktiver erfolgen bzw. die Modellwahl überdacht werden. Die in der Simulation deutlich hervortretenden Effekte werden im realen Einsatz aber vermutlich nur begrenzte Wirkung entfalten.

In einer Rasch-konformen Simulation konnte gezeigt werden, dass der wahre Wert nicht notwendig durch den im Rahmen der Messung bestmöglichen Personenparameter repräsentiert wird. Das gilt insbesondere auch nicht im Mittel. In großen Randbereichen ist dies sogar regelhaft eher nicht der Fall, wie gezeigt werden konnte. Die Verschiebungen sind an den Rändern größer und fallen im Allgemeinen deshalb weniger ins Gewicht, weil Personenparameter (insbesondere in Randbereichen) in der Praxis nie mit einer solchen Präzision interpretiert werden. Dieser Effekt ist aber über die Skala unterschiedlich und muss additiv zum großen Konfidenzintervall verstanden werden.

Das Guttman-Pattern ist beim praktischen Testen eine Ausnahme, allerdings eine mit besonderer Bedeutung für die Psychometrie. Herausgehoben ist seine Bedeutung überdies beim Standard-Setting. Mit Hilfe des Guttman-Pattern wird hier die Messskala mit der fachinhaltlichen Interpretation der Kompetenzstufenmodelle argumentativ verknüpft. In der Arbeit

[1] Der Infit-Parameter zeigt die Rasch-Konformität eines Items an und liegt optimal bei 1. Abweichungen von 0,1 in beide Richtungen gelten als zulässig, bei einigen Autoren auch mehr.

wird allerdings gezeigt, dass eine Rasch-Skala gegenüber Guttman-Pattern nicht unverzerrt ist, dass das der Argumentation zugrunde liegende Interpretationsmuster also gerade nicht von der Empirie gestützt wird.

Die Vorstellung, dass eine Auswahl von Items aus einem bestimmten Schwierigkeitsbereich der Messskala, eine Messung in genau diesem Skalenbereich verbessert, ist so nicht zutreffend. Tatsächlich befindet sich der Bereich der größten Präzision einer Messung dort, wo der Mittelwert der Schwierigkeitsparameter der Items liegt. Die Lage möglicher Personenparametern ist sehr robust gegenüber Veränderungen der Lage einzelner Items auf der Skala. Das bedeutet in der Konsequenz: Dort, wo ein Testinstrument über die Lage von Itemparametern für die Auswahl durch die Schulpraxis spezifiziert wird, wird ggf. ein falscher Eindruck erzeugt. Wenn ein Testheft zum Beispiel durch einige besonders schwierige Items für eine besonders leistungsfähige Klientel angepasst werden soll, ist nicht die konkrete Lage der schwierigen Items für eine ausgeprägte Differenzierung relevant, sondern der sich dadurch verschiebende Mittelwert der Itemschwierigkeiten. Der Effekt könnte also durch die Auslassung von sehr einfachen Items identisch erzeugt werden. Für die Interpretation, wo ein Testheft besonders präzise misst, ergo besonders gut differenziert, ist ein Blick auf die möglichen Personenparameter von größerer Relevanz. Eine hilfreiche Empfehlung beruht daher vermutlich eher auf der Deklaration mit wie viel korrekt gelösten Items, welche Stufe erreichen werden kann. Vielleicht aber ist eine Testheftempfehlung durch eine inhaltliche Spezifikation noch deutlich unterrichtsnäher.

7.1.2. Generalisierung

Des Weiteren wurden längsschnittliche Interpretationen von Ergebnissen aus Vergleichsarbeiten einer Prüfung unterzogen. Die Politik nutzt solche Interpretationen und unterstützt damit eine äquivalente Lesart der Öffentlichkeit. Die Schulen suchen solche Möglichkeiten und irritieren sich bei der Gegenüberstellung von Ergebnissen aus VERA-8, und solchen aus den Prüfungen zum mittleren Schulabschluss oder denen zum Abitur. Nicht zuletzt bedient sich auch die Wissenschaft solcher Exemplifikationen.

Die KMK-Papiere aus den ersten Jahren der VERA-Entwicklung insinuieren solche Interpretationen mit den Ergebnissen der Vergleichsarbeiten und den Erhebungen zum Bildungstrend und nicht zuletzt legt die Verwendung der identischen Skala mit den identischen Kompetenz(stufen)modelle solcherlei Interpretation nicht nur nahe sondern fordert sie heraus. In einem frühen Papier nahm das IQB gar Stellung (Pant et al., 2012) zu einer Interpretation

von Ergebnissen aus VERA-3 2010 und aus dem Ländervergleich 2011 und führt Differenzen auf verschiedene Ursachen zurück, stellt dabei aber die grundsätzliche Interpretation nicht in Frage.

Vorliegende Untersuchungen lassen allerdings Zweifel aufkommen, ob solche längsschnittlichen Interpretationen durch die empirischen Daten gedeckt werden. Konkret wurde festgestellt:

Während sich einzelne Kompetenzen bei der jährlichen Messung im Rahmen der Vergleichsarbeiten auf Landesebene als instabil erweisen, zeigen die gleichen Kompetenzen bei der Messung im Bildungstrend kaum oder deutlich kleinere Veränderungen. Das verwundert, denn bei den nur alle 5 oder 6 Jahre stattfindenden Erhebungen zum Bildungstrend hätte man die größeren Veränderungen vermutet. Unterstellt man die Korrektheit der Messungen bei den Vergleichsarbeiten, könnten die Ergebnisse des Bildungstrends trotzdem zutreffend sein. Sie wären dann aber Lernstandsfeststellungen, die einen mehr oder weniger zufälligen Wert aus stark heterogenen Kohorten zeigen. Die beim Bildungstrend berichteten Vertrauensintervalle unterstützen diese Deutung genauso wenig, wie die im Allgemeinen über die Jahre stabilen Kompetenzstufenverteilungen.

Markant sind hingegen die sehr ähnlichen Veränderungen zwischen den zwei Ländern Berlin und Brandenburg. Insbesondere fällt auf, dass sich der von den VERA-Schwankungen nahezu unbeeinflusst scheinende, über die Jahre verringernde Leistungsabstand zwischen Berlin und Brandenburg in den Ergebnissen zum Bildungstrend in vergleichbaren Größenordnungen widerspiegelt. Das Papier vom IQB zur Vergleichbarkeit zwischen Ergebnissen des Ländervergleichs und der Vergleichsarbeiten (Pant et al., 2012) behandelt demnach kein einmaliges Problem; Folgeuntersuchungen stehen weiterhin aus. Bezüglich der Kompetenzstufen finden sich Artefakte, die insbesondere eine längsschnittliche Interpretation nur vor einem Hintergrund erlauben, der über diese möglichen Artefakte und ihre Größenordnung informiert. Die Abbildung 5.6 (unten) auf Seite 130 zeigt eine von äußeren Einflüssen vollkommen unabhängige Simulation, bei der auf der Basis der verschiedenen VERA-Testhefte eine identische Population Kompetenzstufenverteilungen produziert, deren Unterschiedlichkeit unerwartet groß ist. Hier zeigen sich Unterschiede von Ausprägungen, die auf keine realen Verhältnisse verweisen, sondern nur auf prozessintern erzeugte Artefakte. Auch hier finden sich, wenn auch selten, Schwankungen von 10% zwischen zwei Jahren. Vor jeder (!) Interpretation solcher Unterschiede müssen die Zusammenhänge genau geprüft werden. Der vermeintlich einfache Vergleich von zwei Kompetenzstufenverteilung erweist sich damit als potentieller Fehlschluss.

Nicht zuletzt zeigen die zwei Studien, die mit VERA-Instrumente operieren, dass eine Veränderungsmessung mit einem identischen VERA-Instrument zu zwei Zeitpunkten auch mit kleinen Stichproben zu sehr plausiblen Ergebnissen führen kann. Indes zeigen sich zwischen den Messungen mit zwei verschiedene VERA-Instrumente erhebliche und unerklärbare Differenzen, die jede Interpretation arbiträr erscheinen lassen. Diese Befunde nähren die Vermutung, dass das Linking der VERA-Instrumente über die Jahre nur unzureichend funktioniert. Die Konsequenzen eines tatsächlich nicht funktionierenden Linkings wären allerdings weitreichend. Mit dem aktuellen Instrumentarium der Vergleichsarbeiten wäre bei regelkonformer Nutzung keine valide längsschnittliche Interpretation möglich. Einzige Ausnahme bliebe die Nutzung von identischen Instrumenten oder solchen, die direkt miteinander verlinkt sind. Jede Interpretation von Ergebnissen regulärer VERA-Durchgänge über mehr als einen Zeitpunkt hinaus ohne, dass irgendein Ausgleich vorgenommen wird, müsste abgelehnt werden. Das ursprüngliche Versprechen solcher Nutzung wäre damit nicht eingelöst.

Dies hätte allerdings auch Auswirkungen auf die jährlichen Einzelmessungen, denn jede Messung wird durch die direkte Anbindung der Kompetenzstufen als Ergebnis auf der gleichen Metrik der Bildungsstandards interpretiert. Wenn die Metriken der einzelnen Messungen aber nicht aufeinander beziehbar, also nicht stabil miteinander verlinkt wären, dann müsste die kriteriale Interpretation der VERA-Ergebnisse als nicht valide abgelehnt werden. Diese Interpretation ist aber ein zentrales Element der Vergleichsarbeiten.

Auf der Suche nach einem, die vorgestellten Artefakte zusammenfassendem Befund, findet man bei Maritzen (2014, S. 404):

> „Die Kalibrierung und Normierung der VERA-Tests erfolgt unter Bedingungen, die im Anschluss an den flächendeckenden Einsatz der Tests schwer händelbare Varianzen in den Befunden produzieren (unplausible Varianzen über die Zeit, zwischen den Domänen bzw. Subdomänen etc.).“

und man kann nur vermuten, dass die Kolleginnen und Kollegen in Hamburg beim Versuch, die eigenen KERMIT-Instrumente und die der Vergleichsarbeiten auf einer gemeinsamen Metrik abzubilden, um den Schulen eine Rückmeldung über einen echten Lernzuwachs geben zu können, ähnliche Artefakte aufgefallen sind. Wenn Richter et al. (2014, S.5) zu den „groß angelegten Pilotierungsstudien“ schreiben:

> „Auf der Basis dieser gemeinsamen Datenerhebung gelingt eine psychometrische Anbindung von VERA an die Metrik der Bildungsstandards.“,

bleibt unklar, ob dieses Gelingen des Linkings zwischen Normierung und VERA konkret belegt worden ist oder ob hier nur ein im Erfolgsfall gelingendes Verfahren beschrieben wird. Die vorgelegten Untersuchungen nähren eher Zweifel daran, dass die von der KMK 2006 (KMK, 2006b) vorgeschlagene psychometrische Anbindung von VERA an die Messungen zum Bildungstrend über die gemeinsame Metrik der Bildungsstandards funktioniert.

Diese Problemlage ist nicht neu, wenn auch selten vorgetragen. Das liegt wohl auch daran, dass einzelne Länder eigene Prozeduren entwickelt haben, mit denen die Ergebnisse der Rückmeldungen plausibilisiert werden. Schon 2011 berichtete Kuhn (2011) von ähnlichen Verwerfungen, wie sie hier dargelegt wurden und resümiert, dass „die sog. ‚Frühwarnfunktion' von Vera Ergebnissen sowohl auf der Ebene der Einzelschule als auch systemisch für ein Land wenig überzeugend [ist], da die Bezugnahme auf die ursprünglichen Normierungsstudien des IQB z.T. keine sinnvolle Interpretation zulässt". Auch sieht er die Diskrepanz zwischen der Erwartung, dass „die Ergebnisse eines Landes sich nicht in großen Sprüngen" sondern moderat ändern und den tatsächlichen Veränderungen bei den Vergleichsarbeiten und reklamiert Analysen.

Eine Klärung der Gründe für die dargelegten Befunde ist unbedingt notwendig. Für die Zeit bis zu einer umfassenden Klärung und vielleicht auch danach, wurde ein pragmatischer Lösungsansatz vorgeschlagen. Dieser unterstellt drei Voraussetzungen als gegeben:

1. Die Landesergebnisse der Messungen beim Bildungstrend sind valide.

2. Die reale Veränderung der Kompetenzstände zwischen zwei Jahren sind für ein Land so gering, dass die Annahme, dass keine Veränderung vorliegt, einen nur vernachlässigbar kleinen Fehler produziert.

3. Die Feststellung in dieser Arbeit, dass ein VERA-Instrument eine für sich genommen reliable Messung darstellt, die bezogen auf die Skala der Bildungsstandards nur durch einen unbekannt großen Parameter (linear) verschoben ist, wird akzeptiert.

Die Ergebnisse der Vergleichsarbeiten können dann jedes Jahr so transformiert werden, dass der Mittelwert (und ggf. auch die Standardabweichung) dem des letzten Bildungstrends entspricht, wobei für den zeitlichen Versatz von einem Schuljahr eine entsprechende Korrektur vorgenommen werden muss. Die Annahme ist, dass der mit dieser Prozedur verursachte Fehler im Verhältnis zur aktuellen Verschiebung klein ist. Zudem kann er von der Größe her zumindest grob abgeschätzt werden. Die im Bildungstrend festgestellten kleinen tatsächlichen Veränderungen über mehrere Jahre und die deutlich größeren Unterschiede zwischen

den VERA-Messungen unterstützen ein solches Vorgehen. In der Folge wären alle gemessenen Werte, der Vergleichsarbeiten sowie der Bildungstrends, plausibel miteinander verbunden. Die Unterstellung, dass alle Werte einer gemeinsamen Skala entstammen ist kommunizierbar. Die Bezüge zu den in den Kompetenzstufenmodellen berichteten kriterialen Normen sind weitgehend korrekt, können zumindest als stabil angenommen werden.

Als Ergebnis der Aufklärung der Gründe für die problematischen Befunde könnte sich ergeben, dass eine hinreichende Stabilität nur dann gewährleistet werden kann, wenn die Prozesse der Testentwicklung bei den Vergleichsarbeiten denen der Entwicklung beim Bildungstrend angeglichen werden. Dies würde enorme Aufwendungen verursachen, die ggf. in keinem angemessenen Verhältnis zum Nutzen stehen. Selbst dann könnte die hier beschriebene Prozedur den besten Weg darstellen, plausible Ergebnisse zu berichten.

7.1.3. Rückmeldungen

Die als Vollerhebung ermittelten Quoten für den Abruf der Rückmeldungen sind ernüchternd. Betrachtet wurde der Zeitraum bis zu den Herbstferien des auf VERA-8 folgenden Schuljahres, wobei schon ab den Sommerferien kaum mehr Aktivitäten zu verzeichnen sind. In die Analyse sind die jeweils ersten Abrufe einer jeden Rückmeldung einbezogen worden. Zwischen 35 und 42% der wenige Minuten nach Abschluss der Eingabe bereitgestellten Sofortrückmeldungen[2] werden direkt nach der Eingabe bzw. im Eingabezeitraum heruntergeladen. Bis zum Ende des betrachteten Zeitraums werden für die Fremdsprachen etwas weniger als 80% der Sofortrückmeldungen heruntergeladen, für Deutsch und Mathematik überschreitet die Abrufquote 60% kaum. Sämtliche andere Rückmeldungen liegen auf diesem Niveau oder darunter. Die für die Unterrichtsentwicklung zentrale klassenbezogene Rückmeldung liegt dabei immer mit ca. 50% am unteren Ende. Die Untersuchung schulformbezogener Abrufquoten zeigt beträchtliche Differenzen in der Form, dass beispielsweise für Mathematik die Abrufe von Gymnasien in etwa doppelt so hoch liegen, wie jene der nicht-gymnasialen Schulformen. Für das Fach Deutsch zeigt sich diese Deutlichkeit nicht für jede Rückmeldung, aber selbst die bei den Fremdsprachen kleinsten schulformbezogenen Unterschiede liegen bei ca. 20%. Der Effekt ist also über sämtliche Fächer so zu finden. Auch die zwei untersuchten Länder unterscheiden sich deutlich: Die Rückmeldungen von Schulen aus Land 1 werden deutlich häufiger heruntergeladen, als ihre Äquivalente im Land 2. Logistische Regressionen für alle drei untersuchten Domänen und den jeweils vier Rückmeldungen konnten zeigen, dass selbst beim Einbezug

[2] Tatsächlich wird ausschließlich die Sofortrückmeldung direkt nach der Eingabe zur Verfügung gestellt, alle anderen erst zu einem späteren Zeitpunkt.

der beiden Merkmale *Schulform* und *Land* nur in einem einzigen Fall das Merkmal *Land* ohne signifikanten Effekt berichtet wurde. Ein durch zwei Gruppen repräsentierter Effekt der Testleistung wurde hingegen von der Variable *Schulform* dominiert.

Dass die Abrufquoten dermaßen niedrig sind, aber auch die erheblichen Unterschiede zwischen den Schulformen und Ländern sind unerwartet. Ansätze, wie dem unzulänglichen Abruf von Rückmeldungen entgegengewirkt werden kann, könnten sich zumindest teilweise aus einer tiefer gehenden Anaylse der vorliegenden Daten ergeben. Insbesondere eine vergleichende Betrachtung der Abrufe des Corona-bedingt in Berlin und Brandenburg freiwilligen VERA-8-Durchgangs von 2021 bietet dafür Potential. Aber auch die Betrachtung von Mehrfachabrufen, von konkreten Abrufzeitpunkten oder von schulspezifischen Abrufprofilen, ggf. mit einer ergänzenden Befragung, versprechen neue Einblicke in die schulische Ergebnisnutzung. So könnte die Dissemination von Rückmeldungen in der Fläche beobachtet werden. Eine Auswahl von Interviewpartner*innen auf der Basis solcher als Vollerhebung vorliegender Daten ermöglichte ein reales Gesamtbild der Ergebnisnutzung und speziell auch der nicht-Nutzung sowie ihrer Gründe. Überdies bietet die in dieser Arbeit erstmals vorgelegte Analyse anderen Forscher*innen methodisches Handwerkszeug für vergleichende Untersuchungen.

Unbeantwortet bleibt allerdings, warum die Lehrkräfte nach der in der Regel verpflichtend auferlegten Vorbereitung, Durchführung, Bewertung und Erfassung der Vergleichsarbeiten so selten die Rückmeldungen auch nur Abrufen. Für die hier im Fokus stehenden, auf Kompetenzstufen bezogenen Rückmeldungen sei angemerkt, dass die Ergebnisse der Rückmeldungen zuerst fachdidaktisch eingeordnet werden müssen. Die Rückmeldung selbst tut dies nur rudimentär (vergleiche die Abbildungen im Anhang A.7.2). Eine Bedeutung erlangen die KMK-Kompetenzstufen für die Schulpraxis ausschließlich im Rahmen der Vergleichsarbeiten. Für Lehrkräfte Berlins und Brandenburgs findet sich im verbindlichen Rahmenlehrplan (LISUM, 2015, S. 11) ein eigenes, auf sämtliche Unterrichtsfächer bezogenes Niveaustufenkonzept. Es stützt sich dabei auf ein Kompetenzmodell, welches sich dort, wo sie von der KMK definiert sind, an die Bildungsstandards anlehnt. Die Niveaustufen des Rahmenlehrplans werden allerdings auf ein Kompetenzentwicklungsmodell bezogen. Die Ergebnisse der Vergleichsarbeiten vor diesem Hintergrund zu interpretieren ist ohne psychometrische Verwerfungen kaum zu bewerkstelligen. Aber auch inhaltlich sind das empirisch evaluierte Kompetenzstufenmodell der KMK und das (lediglich) fachdidaktisch fundierte Kompetenzentwicklungsmodell des Rahmenlehrplans schwierig zu vereinende Konzepte. Dennoch wäre zu prüfen, in wieweit ein direkter Bezug von VERA-Ergebnissen über den Rahmenlehrplan Online (LISUM,

2021) zu konkreten Themen in Schulbüchern oder zu Unterrichtsmaterial Lehrkräften hilfreiche Ansätze bieten kann. Die Zuordnung domänenspezifischer Leistung von Schüler*innen zu Kompetenzstufen ist ein Element. Es ist aber offen, wie dieses Element zu einer Weiterarbeit führen soll. Diedrich (WZB, 2021, 21:29) meint dieses Anschlussproblem, wenn sie zu den Tests aus der Schulpraxis berichtet: „... diese Frage: ‚Jetzt sag mir doch, was ich tun soll!' wird immer noch und immer wieder laut gestellt und am Ende des Tages auch nicht wirklich sinnhaft beantwortet." Projekte wie der Lesecheck Online (https://lesecheck-online.isq-bb.de) des ISQ sind Versuche, die Brücke von einer quantitativ berichtenden zu einer fachdidaktisch aufgeladenen handlungsorientierten Rückmeldung zu schlagen.

7.2. Limitationen

Grundsätzlich ist schon oben die Partialität festgestellt worden, die dem Blick dieser Arbeit auf Validität bei den Vergleichsarbeiten innewohnt. Daher fehlt hier ein globaler Befund und es werden einzelne ergänzende Aspekte zusammengetragen.

Die Überprüfung der Gewissheiten setzt auf Simulationen und macht damit die grundlegenden Eigenschaften der Rasch-Skalierung für eine Untersuchung zugänglich. In die Simulation fließen allerdings Daten aus den VERA-Durchgängen ein. Die Itemparameter sind dabei immer nur eine spezifische Auswahl aus dem Itemuniversum. Zwar legt die Wiederholbarkeit von Effekten mit unterschiedlichen Itemsets aus verschiedenen Fächern bzw. Domänen eine Regelhaftigkeit nahe. Eine mathematisch geschlossene Beschreibung scheint hier teilweise aber möglich, beispielsweise für die Robustheit der Verteilung möglicher Personenparameter gegenüber der Verteilung von Itemschwierigkeiten (nicht regelmäßige Diskretisierung einer rationalen Skala). Wegen der Abwesenheit von in der Praxis üblichen Störgrößen, kann die Relevanz der gezeigten Verzerrungen überschätzt werden. Ein Abgleich der Simulationen mit den Ergebnissen der realen Messungen könnte hier Abhilfe schaffen.

Für die Betrachtung der Stabilität der Ergebnisse von Vergleichsarbeiten sind für die Untersuchungen im Rahmen der üblichen VERA-Tests absichtsvoll Messzeitpunkte ausgewählt worden, die besonders deutliche Artefakte zeigen. Damit werden die Probleme in besonderer Weise betont und können unter Umständen zu einem verzerrten Gesamtbild führen. Wie im vorgehenden Abschnitt schon angemerkt wurde, sind die Erhebungen zu den Projekten VERAMSA und der Untersuchung der Entwicklung der Orthographiekompetenz nicht dazu initiiert worden, um die Stabilität der VERA-Messungen zu untersuchen. VERAMSA greift dabei auf Schüler*innen aus Berliner Gymnasien zurück und lässt deshalb nur eingeschränkte

Schlussfolgerungen zu. Die Größe der Stichprobe ist dann auch unzureichend, um das differenzielle Funktionieren der drei verschiedenen Testhefte zum letzten Messzeitpunkt als signifikant nachzuweisen. Da für den Zuwachs der Orthographiekompetenz große Werte erwartet wurden, konnte hier sogar eine kleine Stichprobe hinreichend sichere Effekte nachweisen. Trotz des Verfahrens, mit dem aus der selbstselektierten Stichprobe eine echte Stichprobe gewonnen werden sollte, kann nicht vollständig belegt werden, dass dies gelungen ist. Eine Wiederholung dieser Untersuchung mit einer größeren, wieder geschichtet gezogenen, aber verpflichtenden Stichprobe wäre hier angezeigt. Trotz dieser Einschränkungen legen die Ergebnisse Vermutungen nahe, wie bestimmte Verwerfungen zwischen verschiedenen Messungen erklärt werden können. Die hier notwendige Absicherung der Schlussfolgerungen durch eine für genau diesen Zweck angelegte Untersuchung steht aus.

Die Limitation der Analyse der Rückmeldeabrufe ist weniger bei den Daten selbst zu suchen. Sie stellen eine Vollerhebung dar und die Datenerfassung ist durch standardisierte technische Prozesse sichergestellt. Allerdings werden die schulischen Prozesse durch die Erfassung der Abrufe allein nur sehr unvollständig abgebildet. So können die Gründe für einen nicht erfolgten Abruf einer Rückmeldung äußerst vielfältig sein und die vorliegenden Daten allein lassen keine eindeutigen Interpretationen zu. Deshalb wurde in der vorliegenden Arbeit im Allgemeinen auf aggregierte Ergebnisse rekurriert. Eine große Chance ergibt sich aber durch die Möglichkeit, die Abrufdaten für eine gezielte Stichprobenziehung beispielsweise für eine Interviewstudie zu nutzen. Konkret könnten dezidiert nicht-Nutzer*innen zu ihren Intentionen befragt werden.

7.3. Fazit

Aus der Sicht des Autors ergeben sich aus den vorliegenden Untersuchungen einige mögliche Optionen für weitergehende Untersuchungen sowie für Entwicklungen der Vergleichsarbeiten.

Die unerwartet großen Differenzen zwischen zwei VERA-Messungen müssen untersucht werden. Erfolgversprechend scheint hier eine nähere Untersuchung des aktuell etablierten Linking-Prozesses. Die Differenzen sind nicht allein relevant für längsschnittliche Interpretation, sondern ebenso für die Interpretation aller Ergebnisse mit Bezug zur BiSta-Metrik und den Kompetenzstufenmodellen.

Die vorgeschlagene Interimslösung würde die Verwerfungen offensichtlich verringern. Vielleicht ergibt sich diese Umsetzung sogar als verhältnismäßig beste Lösung, um schnell konsistente Ergebnisse kommunizieren zu können.

Parallel könnte geprüft werden, ob in Anlehnung an die Veränderungen bei PISA der Übergang zu einer 2PL-Modellierung erfolgen sollte. Ggf. könnten bestehende Items unter Beibehaltung des Schwierigkeitsparameters durch eine Diskrimination ergänzt werden. Ein solcher Umstieg könnte im Rahmen der aktuellen Neuformulierung und Neunormierung der Bildungsstandards erwogen werden. Hier kann im Zuge der Digitalisierung eine umfassende Anpassung der Inhalte erwartet werden, wie auch neue Herausforderungen der Item-Modellierung.

Eine grundlegende Neuentwicklung der Rückmeldungen, die sich streng an der intendierten Nutzung orientiert, aktuelle Hinweise zur Gestaltung berücksichtigt (Schliesing, 2017) und die Schulpraxis in die Entwicklung nicht nur einbeziehen, sondern deren Bedarfe in den Vordergrund stellt, ist nicht zuletzt im Sinne der Gewährleistung von Validität unbedingt zu befördern. Zudem muss VERA mit anderen Prozessen der Qualitätsentwicklung sinnvoll und nachvollziehbar verzahnt werden.

Unterrichtsprozesse sind komplex, die Anforderungen an die Lehrkräfte, diese zu managen, sind eminent. Sie sind mit einem Instrument wie einem Kompetenztest allein nie zu lösen. Die Ergebnisse werden immer nur eine Quelle neben anderen sein, die im Zusammenhang zu interpretieren sind. Hinzu kommt, dass fraglos neben den hier fokussierten reflexiv-rekonstruktiven Ansätzen auch routinisierte und unhinterfragte Praxis den Erfolg von Unterricht ausmacht. Letztendlich ist die Entwicklung der durch VERA gemessenen Kompetenzen ein wesentlicher, aber doch eben nur einer neben anderen Aufträgen, welche der Schule auferlegt sind.

Mit der Analyse der Abrufe von Rückmeldungen steht den VERA-administrierenden Einrichtungen ein wichtiges Werkzeug zur Verfügung. Die dabei gewonnene Informationstiefe ist begrenzt, aber es erlaubt die Beobachtung der Aktivitäten aller Rezipienten. Die überraschenden Zahlen haben gezeigt, dass hier erhebliche Varianzen aufzuklären sind. Die Verknüpfung der Abrufdaten mit den Ergebnissen bieten vielfältige Ansätze für eine Analyse von Faktoren, die eine VERA-Nutzung begünstigen. Konkret kann die Dissemination experimenteller Rückmeldeformen genauso beobachtet werden wie der Erfolg von kommunikativen Maßnahmen im Rahmen der Vergleichsarbeiten. Für eine Neuaufstellung der Vergleichsarbeiten könnte dieses Element sehr nützlich eingesetzt werden.

Dieses Instrument ist aber nicht nur einfach, es ist auch sehr einfach zu korrumpieren. Solche Daten sollten deshalb der Nutzung durch eine die Schule kontrollierende Instanz vorenthalten werden. Ein „Nachsteuern“ der Schulen ist die erwartbare Reaktion und der Wert dieses wertvollen Indikators verloren.

Literaturverzeichnis

Altrichter, H., Moosbrugger, R. & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In K. Maag Merki & H. Altrichter (Hrsg.), *Handbuch neue Steuerung im Schulsystem* (2. Aufl., S. 235–277). Springer VS.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (Hrsg.). (2014). *Standards for Educational and Psychological Testing.* American Psychological Association.

Aneis, V., Mahler, N., Henschel, S. & Krüger, S. (2018). *Vergleichsarbeiten 2019, 8. Jahrgangsstufe Mathematik – Technischer Bericht.* Institut für Qualitätsentwicklung im Bildungswesen. Berlin.

Aneis, V., Mahler, N., Henschel, S. & Krüger, S. (2020). *Vergleichsarbeiten 2020, 8. Jahrgangsstufe Mathematik – Technischer Bericht.* Institut für Qualitätsentwicklung im Bildungswesen. Berlin.

Bach, A., Wurster, S., Thillmann, K., Pant, H. A. & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung - Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft*, *17*(1), 61–84.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K. J. & Weiß, M. (Hrsg.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Leske & Budrich.

Baumert, J. & Lehmann, R. (1997). *TIMSS — Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich - Deskriptive Befunde.* Leske + Budrich. https://link.springer.com/content/pdf/10.1007%2F978-3-322-95096-3.pdf

Becker, B., Weinrich, S., Mahler, N. & Sachse, K. (2019). Testdesign und Auswertung des IQB-Bildungstrends 2018: Technische Grundlagen. *IQB-Bildungstrend 2018 – Mathematische und naturwissenschaftliche Kompetenz am Ende der Sekundarstufe I im zweiten Ländervergleich.* Waxmann.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikationen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* (S. 357–375). Beltz.

Bellmann, J. & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. Theoretische Konzeptualisierung und Erklärungsmodelle. *Zeitschrift für Pädagogik*, *55*(2), 286–308. https://doi.org/10.25656/01:4251

Benzig, M., Brändle, T., Klitsche, S., Lücken, M., Musekamp, F. & Thonke, F. (2016). *KERMIT Kompetenzen ermitteln - Hinweise und Anregungen zur Nutzung von KERMIT für die Unterrichts- und Schulentwicklung.*

Best, H. & Wolf, C. (2012). Modellvergleich und Ergebnisinterpretation in Logit- und Probit-Regressionen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *64*(2), 377–395. https://doi.org/10.1007/s11577-012-0167-4

Bildungsklick. (2005). *Ahnen: Auswertung von VERA 2004 betont Notwendigkeit von Leseförderung.* Verfügbar 16. Juli 2020 unter https://bildungsklick.de/bundeslaender/detail/ahnen-auswertung-von-vera-2004-betont-notwendigkeit-von-lesefoerderung

Borsboom, D., Mellenbergh, G. J. & Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Bremerich-Vos, A. & Böhme, K. (2009). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* (S. 219–249). Beltz.

Bremerich-Vos, A. & Krelle, M. (2017). *Vergleichsarbeiten 2017: 3. Jahrgangstsufe (VERA-3) Deutsch - Didaktische Handreichung - Modul B, Didaktische Erläuterungen Rechtschreibung.* Institut zur Qualitätsentwicklung im Bildungswesen. Berlin.

Chapelle, C., Enright, M. & Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13.

Charité - Universitätsmedizin Berlin. (2021). *Progress Test Medizin: Feedback für Studierende und Einrichtungen.* Progress Test Medizin der Charité. Verfügbar 31. Januar 2021 unter https://progress-test-medizin.charite.de/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Bd. 2nd). Erlbaum.

Deci, E. L. & Ryan, R. M. (2000). The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268.

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

Drucksache, 18/13161. (2018, 9. Januar). Schriftliche Anfrage des Abgeordneten Joschka Langenbrinck zum Thema: Ergebnisse des Vergleichstests der Drittklässler/innen (VERA 3) 2017 II. https://kleineanfragen.de/berlin/18/13161

Eiben, M. & Preuße, D. (2017). *Gesamtauswertung der Schulvisitationen der zweiten Runde (2011–2016) in Brandenburg*. Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Emmrich, R. & Harych, P. (2009). *Vergleichsarbeiten in der Jahrgangsstufe 8 im Schuljahr 2007/2008*. Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Feinberg, R. A. & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Frey, A. (2014). *Validität: Internationaler Forschungsstand Und Umsetzung in Deutschland* (Vortrag). Frankfurt am Main.

GEW Berlin. (2013). *Entlastung für Lehrkräfte bei Vera-8*. GEW - Berlin. Verfügbar 30. Dezember 2020 unter https://www.gew-berlin.de/presse/detailseite/neuigkeiten/entlastung-fuer-lehrkraefte-bei-vera-8/

Graf, T., Emmrich, R., Harych, P. & Brunner, M. (2013). Durchführungseffekte bei Vergleichsarbeiten in Jahrgangsstufe 8. *Empirische Pädagogik*, *27*(4), 459–473.

Graf, T., Harych, P., Wendt, W., Emmrich, R. & Brunner, M. (2016). Wie gut können VERA-8-Testergebnisse den schulischen Erfolg amEnde der Sekundarstufe I vorhersagen? *Zeitschrift für Pädagogische Psychologie*, *30*(4), 201–211. https://doi.org/10.1024/1010-0652/a000182

Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung.* Waxmann.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Harsch, C., Pant, H. A. & Köller, O. (Hrsg.). (2010). *Calibrating standards based assessment tasks for English as a first foreign language - Standard-setting procedures in Germany.* Waxmann.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 143–171). Springer.

Hartig, J., Frey, A. & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 529–545). Springer.

Hartig, J. & Klieme, E. (2006). Kompezenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Springer Medizin Verlag.

Hartung-Beck, V. (2009). *Schulische Organisationsentwicklung und Professionalisierung.* VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91489-3

Harych, P. & Emmrich, R. (2014). *Wie stabil sind Kompetenzmessung bei Vergleichsarbeiten?* (Poster). Frankfurt am Main.

Harych, P. & Emmrich, R. (2019). Zwischen wissenschaftlicher Konstruktion und schulischer Praxis: Nutzung individueller Kompetenzstufenrückmeldungen aus Assessments. In J. Zuber, H. Altrichter & M. Heinrich (Hrsg.), *Bildungsstandards zwischen Politik und schulischem Alltag* (S. 265–286). Springer VS. https://doi.org/10.1007/978-3-658-22241-3_12

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Das Seminar*, (2), 90–112.

Helmke, A. & Hosenfeld, I. (2007a). Beschreibung der Fähigkeitsniveaus 'Schreiben' und 'Lesen – mit Texten und Medien umgehen'.

Helmke, A. & Hosenfeld, I. (2007b). Beschreibung der Fähigkeitsniveaus Mathematik VERA 2007.

Henschel, S. & Stanat, P. (2019). Bildungsstandards als Element der Qualitätssicherung und -entwicklung im deutschen Schulsystem. In E. Kiel, B. Herzig, U. Maier & U. Sandfuchs (Hrsg.), *Handbuch Unterrichten an allgemeinbildenden Schulen* (S. 374–383). Verlag Julius Klinkhardt.

Holz, T., Kellermann, C., Harych, P. & Brunner, M. (2014). *VERA 3: Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2013/14 - Länderbericht Berlin.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Hosenfeld, A. (2010). *Führt Unterrichtsrückmeldung zu Unterrichtsentwicklung? Die Wirkung von videographischer und schriftlicher Rückmeldung bei Lehrkräften der vierten Jahrgangsstufe.* Waxmann.

Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ). (2021). *KERMIT - Qualitätsentwicklung und Evaluation*. KERMIT. Verfügbar 7. Januar 2021 unter https://www.kermit-hamburg.de/

IQB. (2021). *FAQ - Häufig gestellte Fragen*. Verfügbar 20. März 2021 unter https://www.iqb.hu-berlin.de/vera/faq

ISQ. (2021). *Aufgabenbrowser*. Kommentierte Aufgaben zur Diagnose und Förderung auf jedem Kompetenzniveau. Verfügbar 15. August 2021 unter www.aufgabenbrowser.de

Itzlinger-Bruneforth, U., Kuhn, J.-T. & Kiefer, T. (2016). Testkonstruktion. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R* (1. Aufl., S. 21–50). Facultas.

Jäger, S. (2012). *Rezeption und Nutzung von Diagnose- und Vergleichsarbeiten an Schulen. Eine Interviewstudie mit baden-württembergischen Lehrkräften an Haupt-, Realschulen und Gymnasien*. Pädagogische Hochschule Schwäbisch Gmünd, Bibliothek.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*, 1–73.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bundesministerium für Bildung und Forschung.

KMK (Hrsg.). (2003c). Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10). Verfügbar 8. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-Vereinbarung-Bildungsstandards-MS.pdf

KMK (Hrsg.). (2004a). *Bildungsstandards für die erste Fremdsprache (Englisch / Französisch) für den Mittleren Schulabschluss (Jahrgangsstufe 10)*. Luchterhand.

KMK (Hrsg.). (2004b). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss*. Luchterhand.

KMK (Hrsg.). (2004c). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss*. Luchterhand.

KMK. (2004d, 4. Juni). *Ergebnisse der 306. Plenarsitzung der Kultusministerkonferenz*. Ergebnisse der 306. Plenarsitzung der Kultusministerkonferenz. Verfügbar 20. März 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/ergebnisse-der-306-plenarsitzung-der-kultusministerkonferenz.html

KMK (Hrsg.). (2004e). Vereinbarung über Bildungsstandards für den Hauptschulabschluss (Jahrgangsstufe 9). Verfügbar 24. Juli 2020 unter https://www.kmk.org/fileadmi

n/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Haupt.pdf

KMK (Hrsg.). (2004f). Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10) in den Fächern Biologie, Chemie, Physik. Verfügbar 24. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Mittleren-SA-Bio-Che-Phy.pdf

KMK. (2005a). *Bildungsstandards der Kultusministerkonferenz - Erläuerungen zur Konzeption und Entwicklung*. Wolters Kluwer.

KMK (Hrsg.). (2005b). *Bildungsstandards für die erste Fremdsprache (Englisch / Französisch) für den Hauptschulabschluss (Jahrgangsstufe 9)*. Luchterhand.

KMK (Hrsg.). (2005c). *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss (Jahrgangsstufe 9)*. Carl Link.

KMK (Hrsg.). (2005d). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4)*. Luchterhand.

KMK (Hrsg.). (2005e). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss (Jahrgangsstufe 9)*. Luchterhand.

KMK (Hrsg.). (2005f). *Bildungsstandards im Fach Mathematik für den Primarbereich (Jahrgangsstufe 4)* (1. Aufl.). Luchterhand.

KMK (Hrsg.). (2006b). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. LinkLuchterhand.

KMK (Hrsg.). (2010). *Konzeption der Kultusministerkonferenzzur Nutzung der Bildungsstandardsfür die Unterrichtsentwicklung*. Carl Link.

KMK (Hrsg.). (2012a). Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA) (Beschluss der Kultusministerkonferenz vom 08.03.2012). Verfügbar 15. August 2021 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

KMK (Hrsg.). (2016a). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Wolters Kluwer.

KMK (Hrsg.). (2016c). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung*. Carl Link.

KMK (Hrsg.). (2018b). Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA) (Beschluss der Kultusministerkonferenz vom 08.03.2012 i. d. F. vom 15.03.2018). Ver-

fügbar 15. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

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

Koch, U., Groß Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). Von der Evaluation zur Schul- und Unterrichtsentwicklung - Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zur 67. Tagung der Arbeitsgruppe der Empirischen Bildungsforschung (AEPF)* (S. 187–199). Waxmann.

Kohler, B. & Schrader, F. (2004). Ergebnisrückmeldung und Rezeption: Von der externen Evaluation zur Entwicklung von Schule und Unterricht. *Empirische Pädagogik*, *18*(1), 3–17.

Kohrt, P., Mahler, N. & Henschel, S. (2020). *Vergleichsarbeiten 2020, 3. Jahrgangsstufe, Mathematik - Technischer Bericht.* Institut zur Qualitätsentwicklung im Bildungswesen. Berlin.

Kolen, M. J. & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices* (3. Aufl.). Springer-Verlag. https://doi.org/10.1007/978-1-4939-0317-7

Köller, O. (2008). *Zum Verhältnis von Kompetenzstufen, Normierung der Bildungsstandards und standardisierten Lernstandserhebungen* (Präsentation). Wiesbaden. Verfügbar 15. März 2021 unter https://www.emse-netzwerk.de/uploads/Main/EMSE08_Koeller_Verhaeltnis_von_Kompetenzstufen_Bildungsstandards_und_LSE.pdf

Köller, O., Anders, Y., Becker-Mrotzek, M., Dreyer, R., Maaß, K., Prediger, S. & Thiel, F. (2020). *Empfehlung zur Steigerung der Qualität von Bildung und Unterricht in Berlin - Abschlussbericht der Expertenkommission.* Berlin. Verfügbar 30. Dezember 2020 unter https://www.berlin.de/sen/bjf/service/presse/abschlussbericht_expertenkommission_6-10-2020.pdf

Kuhl, P., Harych, P. & Hoth, K. (2011). *VERA 3: Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2010/2011 - Länderbericht Berlin.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin, Brandenburg.

Kuhn, H.-J. (2011, 18. Oktober). Tischvorlage Steuergruppe Vera am 20.10.2011.

Kühn, S. M. & Drüke-Noe, C. (2013). Qualität und Vergleichbarkeit durch Bildungsstandards und zentrale Prüfungen? Ein bundesweiter vergleich von prüfungsanforderungen im

fach mathematik zum Erwerb des mittleren schulabschlusses. *Zeitschrift fur Padagogik*, *59*(6), 912–932.

Kuper, H. & Diemer, T. (2012). Vergleichsarbeiten: Theoretische und empirische Betrachtungen zum Nutzen des Vergleichens. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung – Empirische Befunde und forschungsmethodische Implikationen* (S. 225–245). Springer VS.

Landtag Rheinland-Pfalz, 23. Wahlperiode. (2002). *Stenografischer Bericht der 23. Sitzung (Plenarprotokoll 14/23)*. Mainz.

Lankes, E., Burgmaier, F., Rudoph-Albert, F., Teubner, M. & Werner, S. (2018). *Bildungsbericht Bayern 2018*. Bayerisches Landesamt für Schule. Gunzenhausen.

Lee, W.-C. & Ban, J.-C. (2009). A Comparison of IRT Linking Procedures. *Applied Measurement in Education*, *23*(1), 23–48. https://doi.org/10.1080/08957340903423537

Lehmann, R., Peek, R. & Gänsfuß, R. (1996). *Aspekte der Lernausgangslage und der Lernentwicklung von Schülerinnen und Schülern, die im Schuljahr 1996/97eine fünfte Klasse an Hamburger Schulen besuchten.Bericht über die Erhebung im September 1996 (LAU 5)*. Behörde für Schule und Berufsbildung. Hamburg. https://bildungsserver.hamburg.de/lau/

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (S. 149–167). Verlag für Sozialwissenschaften.

Lissitz, R. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

LISUM (Hrsg.). (2015). Rahmenlehrplan für die Jahrgangsstufen 1-10 - Teil C: Mathematik. Verfügbar 9. August 2021 unter https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/rahmenlehrplaene/Rahmenlehrplanprojekt/amtliche_Fassung/Teil_C_Mathematik_2015_11_10_WEB.pdf

LISUM. (2021). *Rahmenplan Online 1-10 Mathematik*. Bildungsserver Berlin-Brandenburg. Verfügbar 9. August 2021 unter https://bildungsserver.berlin-brandenburg.de/rlp-online/c-faecher/mathematik

Lorenz, J. H. (2005). Zentrale Lernstandsmessung in der Primarstufe - Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. *Zentralblatt für Didaktik der Mathematik*, *37*(4), 317–323. https://doi.org/10.1007/BF02655818

Luecht, R. & Ackerman, T. A. (2018). A Technical Note on IRT Simulation Studies: Dealing With Truth, Estimates, Observed Data, and Residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65–76. https://doi.org/10.1111/emip.12185

Mahler, N., Schipolowski, S. & Weirich, S. (2019). Anlage, Durchführung und Auswertung des IQB-Bildungstrends 2018. In P. Stanat, S. Schipolowski, N. Mahler, S. Weirich & S. Henschel (Hrsg.), *IQB-Bildungstrend 2018 – Mathematische und naturwissenschaftliche Kompetenz am Ende der Sekundarstufe I im zweiten Ländervergleich.* (S. 99–130). Waxmann.

Maier, U. & Schymala, M. (2011). Reduktion von sozialen Disparitäten durch datenbasierte Schulentwicklung? Voraussetzungen für die Rezeption und Nutzung zentraler Testrückmeldungen in Fach- und Gesamtlehrerkonferenzen. In F. Dietrich, M. Heinrich & N. Thieme (Hrsg.), *Neue Steuerung - alte Ungleichheiten? Steuerung und Entwicklung im Bildiungssystem* (S. 291–303). Waxmann.

Maier, U. (2010). Vergleichsarbeiten im Spannungsfeld zwischen formativer und summativer Leistungsmessung. *Die deutsche Schule*, *102*(1), 60–69.

Maier, U., Metz, K., Bohl, T., Kleinknecht, M. & Schymala, M. (2012). Vergleichsarbeiten Als Instrument Der Datenbasierten Schul- Und Unterrichtsentwicklung in Gymnasien. In A. Wacker (Hrsg.), *Schul- Und Unterrichtsreform Durch Ergebnisorientierte Steuerung* (S. 197–224). VS Verlag für Sozialwissenschaften.

Maritzen, N. (2014). Glanz und Elend der KMK-Strategie zum Bildungsmonitoring. Versuch einer Bilanz und eines Ausblicks. *Die Deutsche Schule*, *106*, 398–413.

MBJS (Hrsg.). (2020). Informationsvorlage für die 35. Sitzung der Landesregierung am 9. Juni 2020. Verfügbar 15. Juli 2020 unter https://mbjs.brandenburg.de/media_fast/6288/infovorlage_mbjs_regelbetriebschulen_2020-06-09_1130uhr_1.pdf

Medizinischer Fakultätentag. (2021). *Nationaler Kompetenzbasierter Lernzielkatalog Medizin.* Verfügbar 31. Januar 2021 unter http://www.nklm.de/kataloge/nklm/lernziel/uebersicht

Mertens, A., Duske, K., Raschke, R., Berger, J., Hoffmann, J., Harych, P. & Georg, W. (2000). Erste Erfahrungen mit der Einführung eines Progress-Tests an einer deutschen Medizinischen Fakultät. *Medizinische Ausbildung - Supplement der Zeitschrift 'Das Gesundheitswesen'*, 126.

Messick, S. (1989a). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (S. 13–103). American Council on Education and National Council on Measurement in Education.

Messick, S. (1989b). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11.

Ministerium für Bildung, Jugend und Sport. (2021). *ZENSOS - Allgemein*. ZENSOS - Allgemein. Verfügbar 20. März 2021 unter https://mbjs.brandenburg.de/sixcms/detail.php/lbm1.c.263581.de

Nachtigall, C. (2005). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2005-landesbericht.pdf

Nachtigall, C. (2008). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00028635/Kompetenztest2008-Landesbericht.pdf

Nachtigall, C. (2010). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2010-landesbericht.pdf

Nachtigall, C. (2011). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2011-landesbericht.pdf

Nachtigall, C. (2012). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2012-landesbericht.pdf

Nachtigall, C. (2013). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2013-landesbericht.pdf

Nachtigall, C. (2014). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2014-landesbericht.pdf

Nachtigall, C. (2015). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2015-landesbericht.pdf

Nachtigall, C. (2016). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2016-landesbericht.pdf

Nachtigall, C. (2017). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2017-landesbericht.pdf

Nachtigall, C. (2018). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2018-landesbericht.pdf

Nachtigall, C. (2019). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2019_landesbericht.pdf

Nachtigall, C. (2020). *Landesbericht - Thüringer Kompetenztest*. Friedrich-Schiller-Universität. Jena. Verfügbar 28. Dezember 2020 unter https://www.kompetenztest.de/download/kt2020_landesbericht.pdf

Nachtigall, C. & Hellrung, K. (2013). Zur zeitlichen Entwicklung der Rezeption von Vergleichsarbeiten. *Empirische Pädagogik*, *27*(4), 423–441.

Naumann, C. (2008). Zur Rechtschreibkompetenz und ihrer Entwicklung. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch* (S. 134–159). Beltz.

Neuser, J. (2002). *Gegenstandskataloge: Ein Instrument zur Spezifizierung des Prüfungsstoffes der schriftlichen Prüfungen* (Vortrag). Mainz. http://www.mft-online.de/files/61_omft2002_omft2002.pdf

Niedersächsisches Kultusministerium. (2019, 30. Januar). *Pressekonferenz zum Start in das 2. Schulhalbjahr 2018/19*. Start ins 2. Schulhalbjahr 2018/2019. https://www.mk.niedersachsen.de/startseite/aktuelles/presseinformationen/start-ins-2-schulhalbjahr-20182019-unterrichtsversorgung-bei-994-prozent-1137-neue-lehrkraefte-eingestellt-manahmenpaket-zur-entlastung-von-lehrkraeften-vorgestellt-173447.html

OECD. (2003). *PISA 2000 Technical Report*. OECD Publishing.

OECD. (2017). *PISA 2015 Technical Report*. OECD Publishing.

Pant, H. (2011). Bildungsstandards und Vergleichsarbeiten als Instrumente der Qualitätssicherung im Bildungswesen. In Friedrich-Ebert-Stiftung (Hrsg.), *Schulentwicklung zwischen Autonomie und Kontrolle - Wie verändern wir Schule wirklich?* (S. 38–40). Friedrich-Ebert-Stiftung.

Pant, H., Stanat, P., Richter, D. & Weirich, S. (2012). Fachliche Stellungnahme zur Vergleichbarkeit zwischen den Ergebnissen des IQB-Ländervergleichs 2011 in der Primarstufe und den Ergebnissen der Vergleichsarbeiten (VERA-3) des Jahres 2010.

Pant, H., Stanat, P., Schroeders, U., Roppelt, A. & Siegele, T. (2013). *IQB-Ländervergleich 2012 - Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Waxmann. Verfügbar 20. März 2021 unter https://www.iqb.hu-berlin.de/bt/lv2012/Bericht/Bericht.pdf

Pant, H., Tiffin-Richards, S. & Köller, O. (2010). Standard-Setting für Kompetenztests in Large-Scale-Assessments. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (S. 175–188). Beltz.

Pant, H., Tiffin-Richards, S. P. & Stanat, P. (2017). Standard Setting: Bridging the Worlds of Policy Making and Research. In S. Blömeke & J. Gustafsson (Hrsg.), *Standard Setting in Education - The Nordic Countries in an International Perspective* (S. 49–68). Springer.

Penk, C., Roppelt, A., Katzenbach, M. & Pant, H. A. (2014). *Vergleichsarbeiten 2015, 8. Jahrgangsstufe Mathematik - Technischer Bericht.* IQB.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (2013). *PISA 2012 - Fortschritte und Herausforderungen in Deutschland.* Waxmann.

Pukrop, J. (2019). *Rückmeldungen aus Schulleistungstests an Lehrkräfte durch interaktive Informationsvisualisierungen.* Staats- und Universitätsbibliothek Bremen.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing.* Manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* The Danish Institute of Educational Research.

Reinecke, J. (2012, 14. Februar). *Strukturgleichungsmodelle in den Sozialwissenschaften.* Walter de Gruyter.

Reiss, K. & Winkelmann, H. (2009). Kompetenzstufenmodelle für das Fach Mathematik im Primarbereich. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* (S. 120–141). Beltz.

Richter, D., Böhme, K., Becker, M., Pant, H. A. & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. Zusammenhänge zu Verän-

derungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Zeitschrift für Pädagogik*, *60*(2), 225–244.

Robitzsch, A., Kiefer, T. & Wu, M. (2020). *TAM: Test Analysis Modules*. Manual. https://CRAN.R-project.org/package=TAM

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., vollst. überarb. und erw. Aufl.). Huber.

Rost, Jürgen (Hrsg.). (2005). *Die Messtheorie von Rasch in Psychologie und Pädagogik - Internationale Fachtagung vom 25. bis 27. November 2004 im IPN an der CAU Kiel [DVD]*. Kiel, Pabst Science Publishers.

Rychen, D. S. & Hersh Salganik, L. (2001). *Defining and selecting key competencies*. Hogrefe & Huber.

Schaper, N. (2014). Validitätsaspekte von Kompetenzmodellen Und -Tests Für Hochschulische Kompetenzdomänen. In F. Musekamp & G. Spöttl (Hrsg.), *Kompetenz Im Studium Und in Der Arbeitswelt. Nationale Und Internationale Ansätze Zur Erfassung von Ingenieurkompetenzen* (S. 21–48). Lang.

Scheerens, J. (2007). The Case of Evaluation and Accountability Provisions in Education as an Area for the Development of Policy Malleable System Level Indicators. In H.-H. Krüger, T. Rauschenbach & U. Sander (Hrsg.), *Bildungs- und Sozialberichterstattung* (S. 207–224). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90615-7_16

Schildkamp, K. & Teddlie, C. (2008). School performance feedback systems in the USA and in the Netherlands: A comparison. *Educational research and evaluation*, *14*(3), 255–282. https://doi.org/10.1080/13803610802048874

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

Schreiner, C., Harych, P. & Wiesner, C. (2020). Kompetenzstufen in Studien zur Kompetenzmessung im Vergleich: Konzepte, Entwicklung und Interpretation. In U. Greiner, F. Hofmann, C. Schreiner & C. Wiesner (Hrsg.), *Bildungsstandards - Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* (S. 388–409). Waxmann.

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

Senatsverwaltung für Bildung, Jugend und Familie. (2010, 1. September). *Berliner Schülerinnen und Schüler im normierten Bundesdurchschnitt / Keine falschen Schlussfolgerungen aus VERA 3*. Verfügbar 19. Juli 2020 unter https://www.berlin.de/rbmskzl/aktuelles/pressemitteilungen/2010/pressemitteilung.55793.php

Senatsverwaltung für Bildung, Jugend und Familie. (2019). *Schulqualität - Maßnahmen*. Verfügbar 7. November 2020 unter https://www.berlin.de/sen/bildung/unterricht/schulqualitaet/massnahmen/

Senatsverwaltung für Bildung, Jugend und Familie. (2021, 18. Februar). *Der Schulvertrag*. Verfügbar 20. März 2021 unter https://www.berlin.de/sen/bildung/unterricht/schulqualitaet/schulvertrag/

Senatsverwaltung für Bildung, Jugend und Sport. (2003, 19. November). *Vergleichsarbeiten in der 4. Klasse der Berliner Grundschule - Berlin startet am 20. November in 21 Grundschulen*. Verfügbar 12. Januar 2021 unter https://www.berlin.de/rbmskzl/aktuelles/pressemitteilungen/2003/pressemitteilung.42095.php

Siegele, T., Schroeders, U. & Roppelt, A. (2013). Anlage und Durchführung des Ländervergleichs. In H. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegele & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012 - Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 101–121). Waxmann.

Stamm, M. (2003). *Evaluation und ihre Folgen für die Bildung - Eine unterschätzte pädagogische Herausforderung*. Waxmann.

Stanat, P., Böhme, K., Schipolowski, S. & Haag, N. (Hrsg.). (2016). *IQB-Bildungstrend 2015 - Sprachliche Kompetenzenam Ende der 9. Jahrgangsstufe im zweiten Ländervergleich*. Waxmann.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Henschel, S. (Hrsg.). (2019a). *IQB-Bildungstrend 2018 - Mathematische und naturwissenschaftlicheKompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich*. Waxmann.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S. & Henschel, S. (Hrsg.). (2019b). *IQB-Bildungstrend 2018 - Mathematische und naturwissenschaftlicheKompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich: Zusatzmaterial*. Waxmann.

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (Hrsg.). (2017a). *IQB- Bildungstrend 2016 - Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufeim zweiten Ländervergleich*. Waxmann.

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (Hrsg.). (2017b). *IQB- Bildungstrend 2016 - Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufeim zweiten Ländervergleich: Zusatzmaterial*. Waxmann.

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 VS. https://doi.org/10.1007/978-3-658-23240-5_1

Thonke, F. & Lücken, M. (2014). *KERMIT - Kompetenzen ermitteln*. Hamburg. https://www.emse-netzwerk.de/uploads/Main/19E_Luecken_Thonke_EMSE_19_KERMIT_HH.pdf

Tiffin-Richards, S. P., Anand Pant, H. & Köller, O. (2013). Setting Standards for English Foreign Language Assessment: Methodology, Validation, and a Degree of Arbitrariness. *Educational Measurement: Issues and Practice*, *32*(2), 15–25. https://doi.org/10.1111/emip.12008

Toulmin, S. E. (2003). *The Uses of Argument - Updated edition*. Cambridge University Press.

Trendtel, M., Pham, G. & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R – Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (1. Aufl., S. 442). facultas.

van den Ham, A.-K. (2015). *Ein Validitätsargument für den Mathematiktest der National Educational Panel Study für die neunte Klassenstufe*. Leuphana Universität. Lüneburg.

Vereinigung der Oberstudiendirektoren des Landes Berlin e.V. (2017). *Umgestaltung der Einführungsphase in die gymnasiale Oberstufe*. Verfügbar 30. Dezember 2020 unter http://oberstudiendirektoren.de/umgestaltung-der-einfuehrungsphase-in-die-gymnasiale-oberstufe/

Vettorazzi, K. & Harych, P. (2019). *VERA 3: Kompetenzentwicklung Rechtschreiben - Projektbericht*. Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Vettorazzi, K., Kellermann, C., Harych, P. & Brunner, M. (2015). *VERA-3 - Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2014/15 - Länderbericht Berlin.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Vettorazzi, K., Kellermann, C., Harych, P. & Brunner, M. (2016). *VERA-3 - Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2014/15 - Länderbericht Brandenburg.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Vettorazzi, K., Schilling, A., Harych, P. & Brunner, M. (2017a). *VERA-3 - Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2015/16 - Länderbericht Berlin.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Vettorazzi, K., Schilling, A., Harych, P. & Brunner, M. (2017b). *VERA-3 - Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2015/16 - Länderbericht Brandenburg.* Institut für Schulqualität der Länder Berlin und Brandenburg. Berlin.

Wacker, A. & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg: Zur Einschätzung von Lehrkräften vor und nach der Implementation. *Zeitschrift für Erziehungswissenschaft*, *15*(4), 683–706.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weinert, F. (2001a). Leistungsmessungen in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (2. Aufl., S. 17–31). Beltz Verlag.

Weirich, S. (2016, 12. Mai). VERA-3 Deutsch (Lesen): Inkonsistente Parameter zwischen Pilotierung (2015) und Durchführung (2016).

Weiß Aparicio, P. (2021). Unterstützung von Schulen bei der Testheftwahl: Entwicklung von Testheftempfehlungen für VERA 8 durch Prädiktion von Lösungshäufigkeiten auf Schulebene basierend auf einer Simulation mit der Vorjahreskohorte [unveröffentlichte Masterarbeit].

Wendt, W., Penk, C., Stroppel, S. & Institut für Schulqualität (ISQ). (2017). Allgemeinbildende Schulen in Berlin Und Brandenburg: Ein Überblick. *ISQ-Bericht Zur Schulqualität - Qualitätssicherungsverfahren, Prozess- Und Ergebnisqualität an Schulen in Berlin Und Brandenburg.*

Wright, B. & Linacre, J. (1994). *Reasonable Mean-Square Fit Values.* Institute for Objective Measurement, Inc. Verfügbar 21. Februar 2021 unter https://www.rasch.org/rmt/rmt83b.htm

WZB. (2021, 18. Juni). *Zu viel versprochen? Grenzen und Potentiale der Outputsteuerung [Podcast]* (Nr.; 7). Verfügbar 9. August 2021 unter https://bildungspolitik.blog.wzb.eu/2021/06/18/zu-viel-versprochen-grenzen-und-potenziale-der-outputsteuerung/

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.

A. Anhang

A.1. Quellenanalyse: Die Vergleichsarbeiten und ihre Zielbestimmung

Die folgende chronologische Beschreibung der Entwicklung der Vergleichsarbeiten mit einem speziellen Fokus auf deren Zielbestimmung erfolgt aus der Sicht der Kultusministerkonferenz als letztendlichem Initiator der Vergleichsarbeiten in ihrer heutigen Gestalt. Deshalb wurden in erster Linie Verlautbarungen der Kultusministerkonferenz selbst analysiert sowie solche Dokumente, die im Auftrag der KMK, zum Beispiel durch das IQB als Institut aller Länder, entstanden sind bzw. die über die Website der KMK verlinkt sind. Dazu wurden

- alle 97 Beschlüsse und Veröffentlichungen im Bereich *Qualitätssicherung und Schule*,
- alle 1.386 Pressemitteilungen zum Thema *Vergleichsarbeit* sowie
- alle über die Website der KMK zum Thema *Vergleichsarbeit* angebotenen bzw. direkt verlinkte Dokumente

untersucht. Am Stichtag dem 31.08.2020 fanden sich in den 97 Beschlüssen und Veröffentlichungen 18, die das Suchwort *Vergleichsarbeit* enthielten. Als eindeutige Dokumente mit analyserelevanten Informationen wurden 4 identifiziert. In sämtlichen Pressemitteilungen wurden 28 mit dem Suchwort *Vergleichsarbeit* extrahiert, von denen 12 mit relevanten Informationen in die Analyse einbezogen wurden. Die anderen erwähnten die Vergleichsarbeiten nur oder bezogen sich inhaltlich auf die 4 schon zuvor gefundenen Dokumente. Auch unter den 267 Fundstellen bei der Suche innerhalb der KMK-Webseiten fanden sich einige der schon ausgewählten Fundstellen erneut. Oft wurden zudem identische Dokumente mehrfach angezeigt, sofern sie das Suchwort mehrfach enthielten. Neben zwei Erwähnungen, die nichts mit den Vergleichsarbeiten zu tun hatten und drei verwaisten Links fanden sich 223 Erwähnungen der Vergleichsarbeiten, ohne zusätzliche analyserelevante Informationen. Von den verbleibenden 41 Fundstellen führten 7 zu neuen Informationen, so dass 23 Dokumente bzw. Fundstellen

der Analyse zugeführt werden konnten. Als einzige Ausnahme wurde zusätzlich ein externes Papier von Helmke & Hosenfeld aufgenommen (veröffentlicht in zwei Teilen Helmke & Hosenfeld, 2003a; Helmke & Hosenfeld, 2003b). Dieses Papier beschreibt die Konfiguration der deutschlandweit ersten Implementation der Vergleichsarbeiten, was seine besondere Stellung begründet. Auch wenn VERA-8 als eigenständige Entwicklung betrachtet werden kann, so ist dieser Start der Vergleichsarbeiten in Klassenstufe 4 essentiell für das gesamte Projekt VERA und die Konfiguration prägend für viele, wenn nicht gar sämtliche spätere Realisierungen. Hier wird die Frage der Zielbestimmung der Vergleichsarbeiten durch die Strukturierung verschiedener Rückmeldungen das erste Mal explizit beantwortet. Darüber hinaus bleiben Konkretisierungen der Zielbestimmungen von VERA durch einzelne Länder außen vor.

24.10.1997 - Konstanzer Beschluss - Grundsätzliche Überlegungen zu Leistungvergleichen innerhalb der Bundesrepublik Deutschland

Quelle: KMK (1997).

Dieser Beschluss gilt als der Startschuss für das datenbasierte Qualitätsmanagement im deutschen Bildungswesen. Hier wird allgemein formuliert, dass die Mitglieder der Kultusministerkonferenz die Erarbeitung von Instrumenten zur Evaluation für eine Qualitätssicherung für notwendig erachten und konkreter noch das regelmäßige landesübergreifende Vergleichsuntersuchungen durchgeführt werden müssen, um die Entwicklung grundlegender Kompetenzen zu verfolgen.

23.03.2001 - Pressemitteilung: Buch Leistungsmessungen in Schulen im Auftrag der Kultusministerkonferenz erschienen

Quellen: KMK (2001) sowie Weinert (Kapitel 5 2001d) und Weinert (Kapitel 23 2001c).

Dieses Buch wird in der Pressemitteilung als von der KMK in Auftrag gegeben beworben und enthält zwei von Franz E. Weitert verantwortete, für die hier fokussierten Aspekte relevante Kapitel. Er differenziert im Kapitel 5 „Schulleistungen – Leistungen der Schule oder der Schüler?“ Schulleistungsvergleiche auf Stichprobenebene, um die Leistungen von Ländern zu vergleichen und solchen auf der Ebene der Einzelschule, die entweder längsschnittlich durchgeführt werden sollten oder mindestens wichtige Einflussvariablen erfassen. Als entscheidend hebt er im Kapitel 23

> „... auch und vor allem die Kommunikation der ermittelten Befunde an Ministerien, Schulen, Lehrern und wenn nötig und möglich auch an Schüler, Eltern und

die interessierte Öffentlichkeit.“ (ebenda S.359)

hervor und führt damit verschiedene Rezipienten auf, die Ziel der Kommunikation der Ergebnisse von Leistungsmessungen sein können. Weinert differenziert im Folgenden sehr deutlich zwischen „der Schuladministration und der Lehrerschaft als den wichtigsten Adressaten“ und der - im weitesten Sinne - Öffentlichkeit. Er hebt dabei das besonders notwendige Vertrauenspotential zwischen der Schuladministration als Auftraggeber und den „Beteiligten und Betroffenen“ hervor. Zur Frage, welche aus einer Studie erhobenen Leistungsinformationen in welcher Form dazu geeignet sind, eine „Grundlage für pädagogische und didaktische Reflexionen“ zu sein, müssen aus seiner Sicht Wissenschaftler*innen und Praktiker*innen „in wechselseitiger Offenheit“ (S.360) erproben.

01.03.2002 - Pressemitteilung: Ergebnisse der 297. Plenarsitzung der Kultusministerkonferenz

Quelle: KMK (2002a).

Die KMK einigte sich als Konsequenz aus dem erwartungswidrig schlechten Abschneiden Deutschlands bei PISA 2000 auf sieben Handlungsfelder und erwähnt in deren fünftem „Maßnahmen zur konsequenten Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule auf der Grundlage von verbindlichen Standards sowie eine ergebnisorientierte Evaluation“ das erste Mal die Durchführung und Auswertung von auch schulübergreifenden Vergleichsarbeiten.

24.05.2002 - Pressemitteilung: Ergebnisse der 298. Plenarsitzung der Kultusministerkonferenz

Quelle: KMK (2002b).

Im Rahmen des Beschlusses zur Verständigung über die Erarbeitung gemeinsamer Standards wurde schon festgelegt:

> „In landesweiten Orientierungs- oder Vergleichsarbeiten überprüfen die Länder in eigener Verantwortung, in welchem Umfang die Standards erreicht werden. Dieses Verfahren dient der Qualitätssicherung und begleitet den Lernprozess. Die Überprüfung soll nicht auf das Ende der schulischen Laufbahn konzentriert sein. Damit soll möglichst vielen Schülerinnen und Schülern ermöglicht werden, durch individuelle Förderung die gesetzten Ziele zu erreichen.“

Schon diese frühe Festlegung birgt einen Zielkonflikt in sich: Die Überprüfung des Erreichens der Standards im Sinne eines Monitorings, müsste als *summative* Feststellung zum Zeitpunkt der geplanten Zielerreichung stattfinden, zu spät allerdings für eine individuelle Förderung der Schülerinnen und Schüler. Deshalb soll die Feststellung früher erfolgen. Die Implementation von Orientierungs- oder Vergleichsarbeiten soll primär als ein prozessbegleitendes, also *formatives* Instrument erfolgen, mit der zusätzlichen Aufgabe, einen Zwischenstand für das Erreichen der Standards zu ermitteln, quasi als eine Art Prognose.

27.06.2002 - Pressemitteilung: Nationale Bildungsstandards

Quelle: KMK (2002d).

> „Um die Einhaltung dieser Bildungsstandards zu überprüfen, sollen in den Ländern landesweit Orientierungs- und Vergleichsarbeiten geschrieben werden. Diese Überprüfungen sollen in der Primarstufe beginnen und auch in den weiterführenden Schulen ab Jahrgangsstufe 5 bzw. 7 durchgeführt werden. Ziel eines solchen Tests muss es nach Ansicht der KMK sein, dass möglichst viele Schülerinnen und Schüler durch gezielte Forderung und Förderung die gesetzten Ziele erreichen."

Die Pressemitteilung unterstreicht die primäre Zielstellung eines formativen Instrumentes für die Schule und konkretisiert dafür mit der Primar- und der Sekundarstufe mindestens zwei Zeitpunkte. Davon abgrenzend wird ausgeführt:

> „Über die landesweiten Tests hinaus plant die KMK auch in ausgewählten Fachbereichen und Jahrgangsstufen regelmäßige bundesweite Vergleichsuntersuchungen im internationalen Kontext (wie PISA oder die Deutsch-Englischen Schülerleistungen International, kurz DESI, sowie die Internationale Grundschul-Leseuntersuchung, IGLU). Die aus den Vergleichsuntersuchungen gewonnenen Daten sollen in die künftig vorgesehene Berichterstattung der Kultusministerkonferenz über Bildung in Deutschland (Nationaler Bildungsbericht) einfließen."

Damit sind drei weitere Instrumente beschrieben. Die Zielstellung der (1) *bundesweiten Vergleichsuntersuchung* ist nicht präzise umrissen, es handelt sich aber, angelehnt an die weiterzuführenden (2) internationalen Studien, um ein Monitoring des Bildungssystems als Ganzem bzw. seiner in Bildungsfragen föderal unabhängigen Untergliederungen (Länder). Diese Ergebnisse fließen in die (3) nationale Bildungsberichterstattung ein.

17./18.10.2002 - Ergebnisse der 299. Plenarsitzung der Kultusministerkonferenz

Quelle: KMK (2002c).

> Darüber hinaus werden die Länder in landesweiten bzw. länderübergreifenden Orientierungs- oder Vergleichsarbeiten überprüfen, in welchem Umfang die vereinbarten Standards tatsächlich erreicht werden. Ziel dieses Verfahrens soll es sein, eine Qualitätssicherung zu gewährleisten, sich darüber länderübergreifend auszutauschen und es den Schülerinnen und Schülern in allen Ländern der Bundesrepublik Deutschland zu ermöglichen, in allen Bildungsgängen über individuelle Förderung die gesetzten Ziele zu erreichen.

Bisher wurden bestehende oder geplante Aktivitäten der *Länder in eigener Verwaltung* lediglich in einen gemeinsamen Rahmen gesetzt. Für einen länderübergreifenden Austausch ist hingegen irgendeine Form der Koordination notwendig, wenngleich unklar bleibt, worin dieser Austausch konkret bestehen soll.

16.01.2003 - Pressemitteilung: Antrittsrede der KMK-Präsidentin – Karin Wolff

Quelle: KMK (2003b).

> Die Einhaltung der Standards muss kontrolliert werden. Standards sind sinnlos ohne Kontrolle, ob sie eingehalten und die gesetzten Ziele erreicht werden. Dazu sind Lernstandserhebungen in allen Ländern erforderlich. Deren Ausarbeitung und Auswertung muss ebenfalls länderübergreifend erfolgen und in die Hände einer unabhängigen wissenschaftlichen Institution gelegt werden.

Der hier neu verwendete Begriff der Lernstandserhebung kann übergeordnet interpretiert werden, denn hier wird nicht präzisiert, auf welcher Ebene die Kontrolle stattfinden soll. Die Aufgabe der Umsetzung dieser Lernstandserhebungen soll einer zentralen Einrichtung übertragen werden. Man kann hier schon die Idee des späteren Bildungstrends erkennen.

> In einigen Ländern werden Vergleichsarbeiten geschrieben und durch zentrale Abschlussprüfungen oder Prüfungen mit einem Anteil landesweit einheitlicher Aufgaben ergänzt. Das halte ich für sinnvoll. Den Schulen, die dabei schlecht abschneiden, muss geholfen werden. Schulaufsicht und Fortbildungseinrichtungen sind gefordert.

Demgegenüber werden Vergleichsarbeiten als landesbezogen auszuwertende Arbeiten beschrieben, deren Ergebnisse aber nicht im Sinne einer internen Evaluation, sondern eher als Schulmonitoring regionales Qualitätsmanagement unterstützen sollen.

18.02.2003 - Pressemitteilung: KMK-Präsidentin zu Bildungsstandards – Vorstellung der Klieme Expertise

Quellen: KMK (2003e) und Klieme et al. (2003).

Die Expertise fundiert die Entwicklung der Bildungsstandards, benennt die Beschreibung von Kompetenzmodellen als Anschlussaufgabe und ist Ausgangspunkt für die Festlegung von Verfahren zur Überprüfung der Implementation der Bildungsstandards. Sie zeigt auf, welche Erfahrungen in anderen Ländern gemacht wurden und bereitet Entscheidungen durch das Aufzeigen von Optionen eher vor, als dass sie Empfehlungen gibt. Es werden drei verschiedenen, für die Bildungspraxis relevante Testverfahren zur Überprüfung der Implementation von Bildungsstandards differenziert.

1. Systemmonitoring (auch Bildungsmonitoring)

Beim Systemmonitoring werden Tests verwendet, „um Aussagen über das Kompetenzniveau von Schülerinnen und Schülern zu machen“ und damit das Erreichen der Bildungsstandards zu kontrollieren sowie „Zusammenhänge mit schulischen wie außerschulischen Bedingungen aufzudecken“. Die Aggregation erfolgt „in der Regel nicht auf der Ebene der Einzelschule“, sondern auf Ebene von Ländern oder Schulformen. Für das Bildungsmonitoring werden einige Freiheitsgrade bei der konkreten Ausgestaltung solcher Studien identifiziert:

- Stichprobe vs. Vollerhebung,
- Teilnahmestatus verpflichtend vs. freiwillig (für Schulen, Lehrkräfte oder Schüler/innen),
- Rhythmus der Erhebung,
- untersuchte Jahrgangsstufe(n),
- untersuchte Fächer sowie
- der Umgang mit den Ergebnissen von Einzelschulen bzw. einzelnen Klassen. (S. 102)

Im letzten Punkt finden sich implizit Aspekte von zusätzlichen, über die oben zitierte Aussage hinaus gehende Zielstellungen für den Fall von Rückmeldungen auf Schulebene (oder

detaillierter), die nicht allein der Schule zur Verfügung stehen. Die Nutzung der Ergebnisse für die Arbeit der Schulaufsicht und zur Schulwahl durch Eltern bei öffentlicher Verfügbarkeit sind als Beispiele benannt. Konkret wird empfohlen die Implementation von nationalen Bildungsstandards durch ein darauf bezogenes nationales Bildungsmonitoring zu ergänzen und mit der Durchführung ggf. auch ein zentrales Institut zu beauftragen. Aus den Daten von Systemmonitoring-Untersuchungen können auch Normen abgeleitet werden, die in späteren Untersuchungen eine normorientierte Interpretation der Ergebnisse, also einen Vergleich der Messwerte mit einer Referenzpopulation erlauben.

2. Schulevaluation

Schulevaluationen beschreiben „Verfahren zur Reflexion der eigenen Praxis“ (S.83) zur Überprüfung des Erreichens der pädagogischen Ziele. Sie erlauben damit „Rückschlüsse auf den Erfolg schulischer Programme oder unterrichtlicher Maßnahmen“ (S.99). Voraussetzung ist die Feststellung einer interessierenden Problemlage, welche dann die konkreten Testinhalte spezifiziert. Die Evaluationen werden von der Schule selbst oder durch Externe initiiert. Die Autoren stellen dann fest:

> „Die im Zusammenhang mit dem Bildungsmonitoring [...] genannten Fragen zur konkreten Ausgestaltung gelten auch für die Schulevaluation. Von entscheidender Bedeutung ist hier, ob eine regelmäßige Evaluation für die Schulen [...] verpflichtend gemacht wird, wer wie mit den Daten umgehen soll und welche Konsequenzen [...] ein für die Schule problematisches Evaluationsergebnis hat.“ (S.106)

Es wird auf Erfahrungen verwiesen, nach denen Schulen für die Gestaltung eines Evaluationsprozesses Unterstützung benötigen, angefangen beim Finden einer Fragestellung bis zur Interpretation der Ergebnisse und der Ableitung von Strategien zur Entwicklung.

3. Individualdiagnostik

Natürlich kann die Untersuchung der Leistung von Schülerinnen und Schülern im Sinne einer Individualdiagnose dazu genutzt werden, gezielte Fördermaßnahmen zu identifizieren.

> „Für diesen Zweck ist es in der Regel sinnvoller, einen kleineren Kompetenzbereich detaillierter zu erfassen, als das gesamte Spektrum eines Systemmonitorings mit relativ wenigen Aufgaben pro Schüler abzudecken.“ (S.83)

Die Expertise stellt gleich zu Beginn die möglichen kleinsten Aggregate für alle drei Untersuchungen gegenüber. Diese Betrachtungen sind essentiell, weil sie die Ebene der Ergebnisinterpretation und Weiterarbeit und damit letztendlich der Zielstellung betreffen.

> „Das Erhebungsdesign einer bundesweiten Monitoringstudie ist im Allgemeinen nicht so ausgelegt, dass es Aussagen auf Individualebene erlaubt. Das kleinste Aggregat, zu dem hinreichend genaue und valide Messergebnisse abgeleitet werden können, ist die Schule oder unter Umständen die Klasse. Dasselbe gilt für schulbezogene Evaluationen." (S.107)

Als Gründe für diese Einschränkungen wird ausgeführt:

- Bildungsstandards decken ein weites Fähigkeitsspektrum ab, dass allein schon aus zeitlichen Gründen nicht in einem Test geprüft werden kann. Für Studien mit höherem Aggregationsniveau als der Einzelperson hilft hier ein Rotationsdesign, bei dem verschiedene Schülerinnen und Schüler unterschiedliche Testteile bearbeiten. Die *Vergleichsbasis für Individualergebnisse* geht dabei allerdings verloren.
- Für valide Schlussfolgerungen auf der Ebene der Einzelschülerin bzw. des Einzelschülers ist der Messfehler zu groß. Die Expertise ergänzt aber:

  > „Andererseits beruhen die standardbezogenen Tests auf Modellen der individuellen Kompetenzentwicklung und sind daher hervorragend geeignet, auch für individualdiagnostische Zwecke eingesetzt zu werden." (S.108)

Bestimmend ist hierfür, ob das Testdesign darauf abzielt (a) die Leistung von Schulklassen bezüglich bestimmter Kriterien zu untersuchen oder (b) den Lernerfolg einzelner Schülerinnen und Schüler auf individuelle Lernvoraussetzungen zurückzuführen.

> „Grundsätzlich ist es wichtig, die Grenzen der individualdiagnostischen Aussagekraft einzelner Testanwendungen zu beachten." (S 108)

4. Fazit

Die Expertise umreißt zuerst Gestaltungsoptionen für ein Bildungsmonitoring zur Überprüfung der Implementation der Bildungsstandards, führt genau diese aber auch für schulische Evaluationen ins Feld und stellt letztendlich fest:

> „So kann man etwa eine Mischung aus Systemmonitoring und Schulevaluation vorsehen, z.B. wenn Schulen an einem landesweiten Evaluationsprogramm teil-

> nehmen, das ihnen Informationen zu ausgewählten, zentral vorgegebenen Qualitätsaspekten gibt. Zu solchen Qualitätsaspekten wird zukünftig sicherlich die Einlösung der nationalen Bildungsstandards gehören.“ (S.83)

Die Expertise plädiert grundsätzlich dafür, dass „Monitoring- und Evaluationsstudien [...] als zwei verschiedene Typen von empirischen Studien angesehen werden.“, auch wenn „streckenweise dieselben Testinstrumente verwendet werden“ (S. 106). Die Unterschiede solcher Studien sind „in der deutschen Öffentlichkeit, aber auch unter Fachleuten bislang unzureichend beachtet worden.“ (S.103). Dies kann als Aufforderung gedeutet werden, die Wahl der unterschiedlichen Gestaltungsoptionen und deren Konsequenzen für die Ergebnisinterpretation beim Design von Studien zum Bildungsmonitoring einerseits und von Schulevaluationen andererseits offensiver zu kommunizieren.

Anfang 2003 - Projektbeschreibung: Vergleichsarbeiten (VERA) - eine Standortbestimmung zur Sicherung schulischer Kompetenzen

Quellen: Helmke und Hosenfeld (2003a) und Helmke und Hosenfeld (2003b).[1]

Das zunächst für 5 Jahre auf Beschluss des Landtages Rheinland-Pfalz am 25.04.2002 beschlossene Projekt VERA, wird hier (wie wortgleich in Helmke, A. (2004): Unterrichtsqualität: Erfassen, Bewerten, Verbessern) vorgestellt und dabei bezüglich der Konzeption und der Zielstellung gegenüber ähnlichen Projekten positioniert. Hierbei werden einige der in der Klieme-Expertise als Freiheitsgrade beschriebene Entscheidungen für die Vergleichsarbeiten anhand der Zielstellungen getroffen:

- Anfang der 4. Jahrgangsstufe,
- jährliche Durchführung (ab 2003 Mathematik, ab 2004 auch Deutsch),
- inhaltliche Orientierung am Rahmenlehrplan (Rheinland-Pfalz),
- Aufgaben zu Mathematik (klassifiziert nach Inhaltsbereichen und Tätigkeitsanforderungen) sowie zu Deutsch, die das Curriculum vollständig abbilden,
- Aufgabenentwicklung durch Expertengruppen,

[1] Der Analyse standen nicht die originalen Artikel zur Verfügung, sondern eine elektronische Version, die sich mit „erschienen in“ auf diese bezieht, mit dem Erstellungsdatum 20. Mai 2003 in unmittelbarer zeitlicher Nähe entstanden ist und von den Autoren mit folgendem Kommentar versehen wurde: „Die hier wiedergegebene Darstzellung ist eine überarbeitete und aktualisierte Version des gleichnamigen Kapitels aus dem Buch ‚Unterrichtsqualität: Erfassung, Bewertung, Verbesserung‘ (Helmke, 2003)“. Seitenzahlen beziehen sich auf diese 12seitige Version.

- Aufgaben werden pilotiert und anschließend normiert.
- Erlauben einen Vergleich mit landesweiten oder länderübergreifenden Normen,
- Keine Rückmeldung von klassen- oder schulbezogenen Ergebnissen an die Schulaufsicht oder das Ministerium, keine Aussagen über die Leistung einer ganzen Region (Land) beabsichtigt.

  „... innerschulische Vergleiche und darauf basierende pädagogische und fachdidaktische Diskussionen [sind] nicht nur möglich, sondern ausdrücklich erwünscht." (S.6)

Wenn auch eine Initiierung der Durchführung von VERA zentral durch die Landesregierung erfolgt, so trägt VERA einige charakteristische Züge, die sie der schulischen Evaluation zuordnen, wie etwa die eigenständige inhaltliche Festlegung von Testinhalten durch die Schule (zumindest für eine Hälfte des Tests) und die Tatsache, dass keine schulbezogene Rückmeldung Adressaten außerhalb der Schule erreicht. Die zentrale Testkonstruktion, die inhaltlich breite Orientierung am Rahmenlehrplan, eine Landesberichterstattung auf der Basis einer Zentralstichprobe und die Einhaltung von Testgütekriterien entspricht dabei eher einem Vorgehen, wie man es von Monitoringstudien kennt. Gleichsam bieten Helmke und Hosenfeld sogar auch Optionen für eine individuelle Analyse an. Die Vergleichsarbeit

> „bietet für jeden Schüler [und deren Eltern] Vergleichsinformationen zum Leistungsstand auf der Klassen-, Schul- und Landesebene und kann so helfen, Über- wie Unterschätzungen des Leistungsniveaus der Kinder in Deutsch und Mathematik zu vermeiden. [...] Insbesondere der Vergleich auf Landesebene stellt einzigartige Informationen bereit, die sehr viel weitergehende Analysemöglichkeiten erlauben als dies auf der Ebene der Einzelschule möglich ist." (S. 6)

Sie warnen aber aus methodischen und inhaltlichen Gründen auch davor, das Ergebnis „als alleinige Entscheidungsgrundlage für die Grundschulempfehlung" zu nutzen und verweisen auf einen „lediglich ergänzenden Charakter" (S.7). VERA wird durch die Universität Landau genau als jene „Mischung aus Systemmonitoring und Schulevaluation" (Klieme et al., 2003, S.83) entwickelt, wie sie die Autoren der Klieme-Expertise als Möglichkeit insinuiert haben.

04.07.2003 - Pressemitteilung: Internationale Vergleichsstudie ausgewählter PISA-Teilnehmerstaaten

Quelle: KMK (2003a).

> „Das Entscheidende am Gelingen der Bildungsstandards ist die Kombination allgemeiner Bildungsziele mit Kompetenzmodellen und Aufgabenstellungen zur Überprüfung“

so Wolff. Die Einhaltung nationaler Bildungsstandards soll demnach landesweit oder länderübergreifend durch entsprechende Tests und Vergleichsarbeiten überprüft werden. Den unterschiedlichen Anforderungen der Schularten soll dabei Rechnung getragen werden. Dieses Vorgehen wird durch eine unabhängige, von den Ländern beauftragte wissenschaftliche Einrichtung gewährleistet werden. Der Präzisierung der Testzeitpunkte auf mindestens zwei, jeweils in der Primar- und der Sekundarstufe, wird für die Sekundarstufe durch den Hinweis auf die Schulformspezifik ergänzt. In der Folge werden die Bildungsstandards jeweils für den Hauptschulabschluss Ende der Klassenstufe 9 und den Mittleren Schulabschluss zum Ende der Jahrgangsstufe 10 differenziert. Weil die Anbindung von Tests mit konkreten Aufgabenstellungen an auf Kompetenzmodelle bezogene Bildungsstandards eine große Herausforderung darstellt, muss die Umsetzung wissenschaftlich abgesichert werden. Auch in anderen Ländern wurden dazu spezifische Institutionen gegründet und Kompetenz aufgebaut.

04.12.2003 - KMK-Beschluss: Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10)

Quelle: KMK (2003d).

An diesem Tag wurden von der KMK die ersten Bildungsstandards beschlossen; es folgten weitere Beschlüsse für andere Zeitpunkte der Schullaufbahn und andere Domänen. Jedes dieser Papiere wird mit folgender oder einer äquivalenten Formulierung eingeleitet:

> „Die Länder verpflichten sich, die Standards zu implementieren und anzuwenden. Dies betrifft insbesondere die Lehrplanarbeit, die Schulentwicklung und die Lehreraus- und -fortbildung. Die Länder kommen überein, weitere Aufgabenbeispiele zu entwickeln und in landesweiten bzw. länderübergreifenden Orientierungs- und Vergleichsarbeiten oder in zentralen oder dezentralen Prüfungen festzustellen, in welchem Umfang die Standards erreicht werden. Diese Feststellung kann zum Abschluss der Jahrgangsstufe 10 erfolgen oder auch schon zu einem früheren Zeitpunkt getroffen werden, um Interventionen zu ermöglichen.“

Unabhängig von Vergleichsarbeiten wird hier die Überprüfung des Erreichens der Standards als Länderaufgabe definiert, die ggf. mit einer zentralen Auswertung der Prüfungen zum Mittleren Schulabschluss erfolgen kann.

> „Die Standards und ihre Einhaltung werden unter Berücksichtigung der Entwicklung in den Fachwissenschaften, in der Fachdidaktik und in der Schulpraxis durch eine von den Ländern gemeinsam beauftragte wissenschaftliche Einrichtung überprüft und auf der Basis validierter Tests weiter entwickelt."

04.06.2004 - Pressemitteilung: Ergebnisse der 306. Plenarsitzung der Kultusministerkonferenz

Quelle: KMK (2004d).

Die Mitteilung verkündet die Gründung des von allen Ländern gemeinsam getragenen „Institut zur Qualitätsentwicklung im Bildungswesen - Wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin" (IQB) und beschreibt als dessen Hauptaufgabe die Überprüfung und Weiterentwicklung der Bildungsstandards und des Weiteren die Stärkung des länderübergreifenden Austauschs zu spezifischen Maßnahmen sowie den

> „Aufbau eines Aufgabenpools zur Standardüberprüfung sowie die Durchführung eines nationalen Bildungsmonitorings. Außerdem unterstützt das IQB die Länder bei der Bildungsberichterstattung über Deutschland."

Mit der Durchführung des nationalen Bildungsmonitorings, also dem heutigen Bildungstrend, wird das IQB hiermit beauftragt. Für das zweite Instrument zur Kontrolle der Standards, die Vergleichsarbeiten, wird hier keine Aussage getroffen.

16.12.2004 - Beschluss: Bildungsstandards der Kultusministerkonferenz - Erläuterungen zu Konzeption und Entwicklung

Quelle: KMK (2005a).

Bei der Beschreibung der Implementation der Bildungsstandards (Punkt 15) wird ausgeführt, dass die Länder verschiedene Bereiche im Fokus haben müssen und führt dabei unter c) Schul- und Unterrichtsentwicklung aus:

> „Die meisten Länder werden, z.B. über ihre Landesinstitute, in Zusammenarbeit mit dem Institut zur Qualitätsentwicklung im Bildungswesen (IQB) die Einhaltung der Standards überprüfen. Die Überprüfung der Standards soll künftig auch bei der Auswertung von Vergleichsarbeiten erfolgen."

Während sich die Rückmeldungen der ersten Umsetzung von Vergleichsarbeiten durch die Universität Landau auf sogenannte Fähigkeitsniveaus beziehen, die sich am Rahmenplan

Rheinland-Pfalz‘ orientieren, wird hier avisiert, dass sich die von der KMK ins Feld geführten Vergleichsarbeiten an den von ihr entwickelten Standards orientieren sollen. Wie die konkrete Zusammenarbeit der Länder mit dem Institut zur Qualitätsentwicklung im Bildungswesen (IQB) bei der Überprüfung der Einhaltung der Standards aussehen soll, wird nicht näher beschrieben.

02.06.2006 - Pressemitteilung: Ergebnisse der 314.Plenarsitzung der Kultusministerkonferenz, Broschüre: Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring

Quellen: KMK (2006a) und KMK (2006c).

Die Überprüfung der Bildungsstandards wird hier erstmals konkret differenziert, in a) Zentrale Überprüfung des Erreichens der Bildungsstandards (jetzt Bildungstrend) mit folgenden Festlegungen

- „Zeitpunkt ca. ein Jahr vor Abschluss des jeweiligen Bildungsgangs“, also in Jahrgangsstufe 3, für den Hauptschulabschluss in Jahrgangsstufe 8 und für den Mittleren Schulabschluss in Jahrgangsstufe 9.
- entsprechend den „üblichen technischen und methodischen Anforderungen“, also als stichprobenbasierte Studie mit Testleitungen.
- Pychometrisch „durch sogenannte Ankeraufgaben (Ankeritems)“ mit den internationalen Studien verknüpft, die im Rahmen dieser Studien normiert werden.
- Für die Lesekompetenz, Mathematik und die Naturwissenschaften.

b) Länderspezifische und länderübergreifende Vergleichsarbeiten in Ankoppelung oder Anlehnung an die Bildungsstandards.

- Vergleichsarbeiten dienen der „Untersuchung des Leistungsstands aller Schulen und Klassen“ (Vollerhebung).
- Durchführung durch die unterrichtenden Lehrkräfte selbst (keine Testleitungen).
- Als Ziel ist beschrieben: „Die Ergebnisse werden in einer kurzen Frist an die Schulen zurückgemeldet, damit sie in die Unterrichts- und Schulentwicklung Eingang finden können.“ Die Ergebnisse sollten „für die gezielte Förderung der untersuchten Klassen genutzt werden“.

Zur Verknüpfung der Vergleichsarbeiten mit dem Bildungsstandards, den vom IQB formulierten Kompetenzstufenmodelle sowie den Ergebnissen aus den Ländervergleichen (also der zentralen Überprüfung nach a), wird angemerkt:

- Durch Verwendung normierter Items soll eine Verknüpfung mit den Bildungsstandards geschaffen werden, wobei zwischen Anlehnung und Ankopplung an die Bildungsstandards differenziert wird. Ankopplung meint eine psychometrische Anbindung der Berichtsskalen. Nur so ist eine Interpretation der Ergebnisse mit Bezug auf die den Ländervergleichen zugrunde liegenden Kompetenzmodellen möglich.
- Normierte Aufgaben dazu liegen für verschiedene Jahrgangsstufen vor, die sich damit für Vergleichsarbeiten empfehlen: „3 und 4 für Deutsch, Mathematik“, „8 und 9 für den Hauptschulabschluss in Deutsch, Mathematik, Erste Fremdsprache (Englisch, Französisch)“ sowie „9 und 10 für den Mittleren Schulabschluss in Deutsch, Mathematik, Erste Fremdsprache (Englisch, Französisch), Biologie, Chemie, Physik.“
- Eine Ankopplung der Vergleichsarbeiten an die Bildungsstandards ermöglicht:
  - Die Durchführung der zentralen Ländervergleiche zur Überprüfung der Bildungsstandards und der Vergleichsarbeiten „in einem nahen Zeitfenster und im selben Jahrgang“,
  - die Verknüpfung der Ergebnisse von Vergleichsarbeiten „direkt mit den Befunden des zentralen Ländervergleichs“,
  - Nutzung der Ergebnisse beider „Untersuchungen für die gezielte Förderung der untersuchten Klassen“.

Das erste Gesamtkonzept beschreibt zudem, dass sich „alle Länder auf ein gemeinsames Verfahren für Vergleichsarbeiten am Ende der Jahrgangsstufe 3 (bzw. Jahrgangsstufe 4 in Brandenburg und Berlin aufgrund der sechsjährigen Grundschule) verständigt“ haben, dass erstmals im Schuljahr 2007/08 durchgeführt wird.

18.09.2006 - Pressemitteilung: Länderübergreifendes Kooperationsprojekt Lernstandserhebungen in der Grundschule geht an den Start

Quelle: KMK (2006d).

Nach der Verabschiedung der Gesamtstrategie zum Qualitätsmanagement, die auch länderübergreifende Vergleichsarbeiten umfasst, schließen sich alle 16 Länder ab 2008 dem Projekt an.

19.10.2006 - Gemeinsame Erklärung mit den Bildungs- und Lehrergewerkschaften

Quelle: KMK (2006d).

In einer gemeinsamen Erklärung von Bildungs- und Lehrergewerkschaften und der Kultusministerkonferenz mit dem Titel *Fördern und Fordern - eine Herausforderung für Bildungspolitik, Eltern, Schule und Lehrkräfte* wird für die Vergleichsarbeiten festgestellt, dass Lehrkräfte

> „... durch flächendeckende Vergleichsarbeiten, die an den länderübergreifenden Bildungsstandards orientiert sind, wichtige Rückmeldungen über den Erfolg ihrer Arbeit und notwendige Verbesserungsmaßnahmen [erhalten]. Diese Vergleichsarbeiten stellen eine klare Beziehung zwischen Bildungsstandards sowie deren Überprüfung und Aspekten des Förderns und Forderns sowie der Unterrichts- und Schulentwicklung her."

Wie schon in der ersten Realisierung der Vergleichsarbeiten durch Helmke und Hosenfeld soll das gemeinsame Projekt VERA über die bloße summarische Feststellung des Erreichens der Standards Informationen liefern, die in eine formative Unterrichtsadaption münden.

18.10.2007 - Pressemitteilung: Ergebnisse der 319.Plenarsitzung der Kultusministerkonferenz

Quelle: KMK (2007).

Im Rahmen der Standardentwicklung für die Abiturphase wurde das IQB beauftragt,

> „ein Konzept einschließlich eines Kostenplans und einen möglichen Zeitplan für die Implementierung von länderübergreifenden Vergleichsarbeiten im ersten Jahr der Qualifikationsphase der gymnasialen Oberstufe zu erstellen."

Da sich später keine weiteren Hinweise auf das Projekt finden, wurde es offenbar verworfen.

10.12.2009 - Broschüre: Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung

Quelle: KMK (2010).

Die KMK erläutert in dem Beschluss eine Konzeption zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung, welche eine Überprüfungs- und eine Entwicklungsfunktion

„in systematischer Weise miteinander“ verbindet. Diese Verbindung manifestiert sich in der Darstellung eines datengestützten Entwicklungskreislaufs, der anzeigt, wie die Einzelschule Prozesse implementieren soll, die „vom Messen zum Entwickeln“ führen. Es wird ausgeführt:

> „Lernstandserhebungen wie z.B. VERA oder andere auf die Standards bezogene und normierte Testinstrumente können, gekoppelt mit solchen Kompetenzstufenmodellen, Schulen Hinweise geben, wie sich Klassen und - in einem sehr eingeschränkten Maße - die einzelnen Schülerinnen und Schüler auf die Niveaustufen verteilen und damit potenziellen Förderbedarf anzeigen.“

Zwar bilden „differenzierte und präzise Befunde von Lernstandserhebungen [...] für sich allein [...] keine hinreichende Voraussetzung für eine gelingende Qualitätsentwicklung“, aber sie sind Ausgangspunkt einer idealtypischen „Verzahnung der Ergebnisse von Lernstandserhebungen mit der Entwicklung von Unterricht“. Die Vergleichsarbeiten werden mit diesem Beschluss deutlich als Instrument der schulischen Evaluation positioniert.

28.04.2010 - Erster VERA-Testtag in Durchführung des IQB

Quelle: Bremerich-Vos et al. (2010).

Das IQB übernahm die Testdurchführung von der Universität Koblenz-Landau. Parallel dazu wurden die vormaligen Fähigkeitsstufen durch Kompetenzstufen ersetzt, die auf Beschlüssen der KMK basieren und die Metrik der Bildungsstandards begründen. Diese Metrik ist auch Grundlage der Berichterstattung des Ländervergleichs, des späteren Bildungstrends. In Didaktischen Handreichungen gibt das IQB Auskunft zur Zielstellung bzw. der Nutzung der Rückmeldung.

> „In diesem Sinne sollen die Vergleichsarbeiten fachliche, fachdidaktische und pädagogisch-psychologische Impulse für die Schul- und Unterrichtsentwicklung bieten. Die aktive Beteiligung der Lehrkräfte an der Durchführung und Auswertung soll zu schulinterner Kooperation und Diskussion bspw. über die Bildungsstandards, die Unterrichtsgestaltung und die eigene Beurteilungspraxis anregen.“ (ebenda, S.4)

Zwar werden die Vergleichsarbeiten als diagnostisches Instrument eingestuft, allerdings auch deutlich gemacht, dass eine regelmäßige Standortbestimmung notwendig ist, um Unterrichtsqualität einschätzen zu können. Als Ziel wird ebenso deklariert, Leistungsunterschiede zwischen Lernenden differenzieren zu können, aber das Ergebnis für eine Einzelschülerin sollte immer im Zusammenhang mit anderen Informationen bewertet werden. Auf Klassenebene

aggregierte Ergebnisse sind hingegen einfacher interpretierbar. Also Ziel wird neben der Einschätzung der Leistungen der Schülerinnen und Schüler bezüglich der Bildungsstandards eine Bewertung des unterrichtlichen Erfolgs genannt, aber auch die Beschäftigung mit kompetenzbezogenen Standards überhaupt.

08.03.2012 - Broschüre: Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten

Quelle: KMK (2012b).

Die KMK fasste 2012 diesen Beschluss über eine „Vereinbarung zur Weiterentwicklung von VERA“, der 2018 erneuert wurde (siehe weiter unten KMK, 2018c). Dieser erste Beschluss fixiert dabei einen Status Quo zwischen allen Ländern bezüglich einiger Aspekte der Zielbestimmung und der Modalitäten der Durchführung und Rückmeldung und zeigt dabei weniger Entwicklungsperspektiven für das Instrument auf. Unmissverständlich wird gleich im ersten Satz des ersten Abschnitts *A) Zielbestimmung von VERA in den Ländern* festgestellt:

> „Die zentrale Funktion von VERA als einem von vier Elementen der Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring liegt in der Unterrichts- und Schulentwicklung jeder einzelnen Schule. Hinzu kommt die wichtige Vermittlungsfunktion, die VERA für die Einführung der zentralen fachlichen und fachdidaktischen Konzepte der Bildungsstandards hat.“

und diese Feststellung wird durch weitere Festlegungen gestützt:

- Eine Veröffentlichung von Ergebnissen einzelner Schulen wird als “mit der Kernfunktion des Instruments, Schul- und Unterrichtsentwicklung zu betreiben, nicht zu vereinbaren“ abgelehnt. Die Identifikation der Einzelschule soll selbst „im Falle kleinräumiger Aufbereitung von VERA-Daten“ unterbleiben.

- Eine individuelle Rückmeldung für Schülerinnen und Schüler bzw. deren Eltern, sei „fachlich vertretbar“, sofern diese „pädagogisch angemessen eingeordnet“ würde, sie muss aber nicht obligatorisch erfolgen. Dass eine Benotung nicht „vorgesehen“ sei, lässt ebenso Spielraum.

- Aber auch „eine Einsicht in VERA-Ergebnisse auf Schul- und Klassenebene“ durch Schulaufsicht und/oder Schulinspektion „kann“ sinnvoll sein, sofern diese damit die Prozesse der Schul- und Unterrichtsentwicklung unterstützen.

Die Vereinbarung schließt mit einer Liste von Maßnahmen, welche auf die Verbesserung der Ergebnisnutzung für die Schul- und Unterrichtsentwicklung zielen. Die Schulleitung soll hierbei Verantwortung für die verbindliche Qualitätssicherung an der Schule übernehmen und VERA als Mittel genau dafür in der Aus- und Fortbildung verankert werden. Eine klare Abgrenzung der Nutzung des Instruments im Sinne eines Monitorings findet nicht statt.

18.04.2013 - VERA 3 und VERA 8: Fragen und Antworten für Schulen und Lehrkräfte

Quelle: KMK (2013).

Dieses Papier wurde von der Amtschefskommission "Qualitätssicherung in Schulen", die von der KMK eingerichtet wurde, um sie bei ihren Vorbereitungen der Plenumssitzungen zu unterstützen, in Auftrag gegeben und zur Kenntnis genommen. Dieses Dokument fasst im Wesentlichen die „für alle Länder geltenden Zielsetzungen, Rahmenbedingungen und Regeln für VERA" zusammen. Neben der Bekräftigung von Schul- und Unterrichtsentwicklung als zentraler Funktion, führt das Dokument auch wieder die Vermittlungsfunktion „für die Einführung der zentralen fachlichen und fachdidaktischen Konzepte der Bildungsstandards" auf. Der formative Charakter der Vergleichsarbeiten wird durch die Einordnung als „Frühwarnsystem für die Unterrichtsgestaltung" kurz vor dem jeweiligen Ende eines Schulabschnitts unterstrichen. Für die Lerngruppe als „wichtigster Analyseebene" werden zwei Bezugsnormen für relationale Vergleiche hervorgehoben:

- der kriteriale Vergleich an den Kompetenzstufen für einzelne Schülerinnen und Schüler sowie
- ein bezugsgruppenorientierter Vergleich, bei dem die Ergebnisse der Lerngruppe denen anderer Lerngruppen und des Landes, ggf. als fairer Vergleich, gegenübergestellt werden können.

Während die Nutzung der Ergebnisse als Rankinginstrument deshalb abgelehnt wird, weil der damit erzeugte Druck dem Ziel einer Weiterentwicklung widerspricht, wird gegen eine „vertiefte Individualdiagnostik" die begrenzte inhaltliche Abdeckung der Tests ins Feld geführt und die Prognosefähigkeit deshalb als unzuverlässig bezeichnet. Bei der Darstellung der Unterschiede zwischen den Vergleichsarbeiten und den Monitoringstudien (international wie national) bleibt hingegen die differente inhaltliche Abdeckung unerwähnt, es werden lediglich die unterschiedlichen Modi bezüglich Stichprobe vs. Vollerhebung und der Durchführung angeführt und auf die verschiedenen Analyseebenen verwiesen.

11.06.2015 - Beschluss: Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring

Quelle: KMK (2016b).

In erste Gesamtstrategie vom 02.06.2006 legte eher basale Strukturen von VERA fest, Aspekte, die später durch die Vereinbarung zur Weiterentwicklung von VERA vom 08.03.2012 (KMK, 2012b) in einem länderübergreifenden Rahmen fixiert wurden. Diese Überarbeitung der Gesamtstrategie benennt mit der *Unterstützung der Unterrichts- und Schulentwicklung* sowie der *Vermittlungsfunktion für die Einführung der fachlichen und fachdidaktischen Konzepte der Bildungsstandards* die bekannten zwei Funktionen als zentral.

Für eine eventuelle Nutzung von Ergebnissen außerhalb der Schule stellt die KMK in dieser überarbeiteten Gesamtstrategie fest, dass die Vergleichsarbeiten *zur Unterstützung der Schulen ggf. von den zuständigen Schulaufsichten oder Schulinspektoraten genutzt werden* können. Darüber hinaus werden die Vergleichsarbeiten in einen Kanon verschiedener Instrumente eingeordnet, *mit denen die Länder eine evidenzbasierte Qualitätsentwicklung und -sicherung auf Ebene der einzelnen Schule gewährleisten* sollen. Dies meint einerseits andere Leistungsvergleichsuntersuchungen, benannt werden hier z.B. Sprachstandsfeststellungen, aber in dieser neuen Gesamtstrategie auch andere Maßnahmen evidenzbasierter Qualitätsentwicklung wie die externe und die interne Evaluationen.

15.03.2018 - Pressemitteilung: Qualitätssicherung: Vergleichsarbeiten werden modernisiert

Quellen: KMK (2018a) und KMK (2018c).

Auch dieser Beschluss bezieht sich auf den vorgängigen Beschluss vom 08.03.2012 über die Vereinbarung zur Weiterentwicklung von VERA (KMK, 2012b), fokussiert nun allerdings eher die Entwicklung des Instruments selbst, die sich damit „noch konsequenter an der Funktion der Unterrichts- und Schulentwicklung ausrichtet“. Der Beschluss stellt präambelhaft fest:

> „Vergleichsarbeiten sind ein pädagogisches Diagnoseinstrument, das Lehrkräften und Schulleitungen zur Verfügung gestellt wird, um festzustellen, über welche fachlichen Kompetenzen Schülerinnen und Schüler einer Lerngruppe verfügen, ...“

Im Folgenden wird der Wert der VERA-Ergebnisse für die „weitere Planung pädagogischer Interventionen und Fördermaßnahmen“ als „Ergänzung zu [...] unterrichtspraktischen Erfahrungen“ der Lehrkräfte hervorgehoben, wie auch deren Rolle bei der Unterstützung des

Diskurses „für eine kooperative Unterrichtsentwicklung im Kollegium“. In diesem Beschluss wird die Weiterentwicklung des Testinstruments VERA in Form einer Flexibilisierung und Modularisierung ausschließlich mit der Erhöhung des schulpraktischen Nutzens begründet.

> “Beide Aspekte dienen [...] einer Stärkung der Verantwortung der Einzelschule für die zielführende Nutzung von VERA hinsichtlich der Auswahl der Fächer bzw. Kompetenzbereiche, der Testgestaltung und der Nutzung der Testergebnisse im Rahmen der Unterrichts- und Schulentwicklung.“

Für die Auswahl aus der sukzessive zu entwickelnden Vielfalt von Modulen wird als Zielstellung formuliert, dass solche „Ergänzungsmodule ausgewählt werden, von denen zu erwarten ist, dass sie dem Kompetenzniveau der Schülerinnen und Schüler besonders gut entsprechen und damit weiteren diagnostischen Erkenntnisgewinn ermöglichen.“. Auch diese Formulierung lässt beträchtlichen Spielraum. Geht es hier darum, durch eine besser Passung eine psychometrisch exaktere Messung zu ermöglichen und damit vielleicht auch eine Nutzung der Ergebnisse auf Ebene der einzelnen Schülerinnen und Schüler? Eine Perspektive, die bisher in öffentlichen Verlautbarungen lediglich als obligatorische Information der Eltern bespielt wurde. Und ist diese präzisere Messung dann die Basis eines weiteren Erkenntnisgewinns? Oder eröffnen sich Lehrkräften auf der Ebene von Rückmeldungen zu inhaltlich optimiert passenden Tests, neue diagnostische Erkenntnisse? Die hier verstärkt zu unterstützende Nutzungsszenarien zielen allesamt ganz klar, auf Unterrichtsentwicklung, im weiteren Sinne auch auf Schulentwicklung.

Auch wenn „eine landesinterne Nutzung von VERA-Daten“ eingeräumt wird, so zitiert der neue Beschluss den alten dahingehend, dass keine Benotung der Vergleichsarbeiten „vorgesehen“ ist, von der Veröffentlichung von Schulergebnissen abzusehen ist, es weder Ländervergleiche noch landesinterne Vergleiche geben soll, wobei dies alles aber auch nicht ausgeschlossen wird. Schulaufsichten soll die Möglichkeit gegebene werden, im Rahmen ihrer beratenden Tätigkeit VERA-Ergebnisse zu thematisieren.

2019 - Das Bildungswesen in der Bundesrepublik Deutschland 2017/2018 - Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa

Quelle: Eckhardt (2019).

In diesem Dossier werden mit der Unterstützung der Schul- und Unterrichtsentwicklung sowie der Vermittlungsfunktion für die Einführung der fachlichen und fachdidaktischen Konzep-

te der Bildungsstandards die zwei zentralen Funktionen benannt und dabei unter *Verfahren zur Qualitätssicherung auf der Ebene der Schule* eingeordnet (S.244).

Tabellarische Zusammenstellung benannter Ziele

Die konkreten Zielstellungen werden hier chronologisch dargestellt und bezüglich dreier verschiedener Wirkungsebene, a) der Mikroebene des Unterrichts, b) der Mesoebene der Schule bzw. c) der Makroebene des Systems zugeordnet. Farblich hervorgehoben sind solche Zielstellungen, bei der außerschulischer Stakeholder Zugang zu Daten innerschulischer Prozesse erlangen.

Eine zusammenfassende Bewertung dieser Quellenanalyse findet sich im Abschnitt 1.3.

Tabelle A.1.: Funktionen der Vergleichsarbeiten aus Sicht der KMK

Quelle	Mikroebene (Klasse)	Mesoebene (Schule)	Makroebene (System)
KMK (2002a)	Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule lernprozessbegleitende Qualitätssicherung, Förderung möglichst vieler Schüler*innen		
KMK (2002b)	Feststellung des Lernstands, Nutzung aber formativ, lernprozessbegleitend		
KMK (2002c)			Erreichen der Standards landesweit und landesübergreifend überprüfen, länderübergreifender Austausch
KMK (2002d)	möglichst viele Schüler*innen sollen gezielte gefordert und gefördert die Standards erreichen		
KMK (2003b)	Einhaltung der Standards kontrollieren	Schulen unterstützen (Schulaufsicht, Fortbildung)	

Fortsetzung auf der nächste Seite

Tabelle A.1.: Fortsetzung der Tabelle von der Vorseite

Quelle	Mikroebene (Klasse)	Mesoebene (Schule)	Makroebene (System)
Helmke und Hosenfeld (2003a)		innerschulische Vergleiche (für Schule und Eltern), pädagogische und fachdidaktische Diskussionen erwünscht	nur anonyme, aggregierte Ergebnisse für die Administration, einzigartige Vergleichsinformation
KMK (2003d)	Implementation der Standards und deren Überprüfung sind Länderaufgabe, die Vergleichsarbeiten ein Instrument der Umsetzung		
KMK (2006b)	Untersuchung des Leistungsstands aller Schulen und Klassen, Unterrichts- und Schulentwicklung Rückmeldung an die Lehrkräfte über den Erfolg der eigenen Arbeit, gezielte Förderung der Klassen mit den Ergebnissen		
KMK (2006d)	VERA soll in formative Unterrichtsadaption münden		
KMK (2010)	Anzeige von potentiellem Förderbedarf einzelner Schüler*innen Beschäftigung mit kompetenzbezogenen Standards	Verbindung der Überprüfungs- und Entwicklungsfunktion, also Etablierung eines datengestützten Kreislaufs für die Einzelschule	
Bremerich-Vos et al. (2010)	Einschätzung von Unterrichtsqualität	Beteiligung der Lehrkräfte führt zu schulinterner Kooperation und Diskussion ...	
KMK (2012a)	Ziel: Unterrichts- und Schulentwicklung		

Fortsetzung auf der nächste Seite

Tabelle A.1.: Fortsetzung der Tabelle von der Vorseite

Quelle	Mikroebene (Klasse)	Mesoebene (Schule)	Makroebene (System)
	Vermittlungsfunktion für fachliche und fachdidaktische Konzepte der Bildungsstandards	Schulaufsicht/-inspektion kann Ergebnisse Schule und Klasse einsehen, soll unterstützen	
KMK (2013)	kriterialer Vergleich für einzelne Schüler*innen bzw. Lerngruppen		Landeswerte als bezugsgruppenorientierter Vergleich
KMK (2016a)	Ziel: Unterrichts- und Schulentwicklung	Nutzung der Ergebnisse durch Schulaufsicht und Schulinspektorate	
	Vermittlungsfunktion für die Einführung der fachlichen und fachdidaktischen Konzepte der Bildungsstandards	eines von verschiedenen Instrumenten schulischer Qualitätsentwicklung und -sicherung	
	Verknüpfung mit interner Evaluation	Verknüpfung mit externer Evaluation	
KMK (2018a) und KMK (2018b)	noch konsequenter Unterrichts- und Schulentwicklung, VERA als pädagogisches Diagnoseinstrument	kooperative Unterrichtsentwicklung im Kollegium	landesinterne Nutzung möglich
	Planung von pädagogischen Interventionen, Fördermaßnahmen	mehr Verantwortung für zielführende Nutzung kommt der Einzelschule zu	
		schulaufsichtliche Nutzung im Rahmen ihrer beratenden Tätigkeit	
Eckhardt (2019)	Ziel: Unterrichts- und Schulentwicklung		
	Vermittlungsfunktion für die Einführung der fachlichen und fachdidaktischen Konzepte der Bildungsstandards		

A.2. Literaturverzeichnis zur Quellenanalyse

Bremerich-Vos, A., Behrens, U., Böhme, K., Engelbert, M., Linkert, D. & Krelle, M. (2010). *Vergleichsarbeiten 2010, 3. Jahrgangsstufe (VERA-3), Deutsch – Didaktische Handreichung zu Testheft II - Rechtschreibung.* Institut für Qualitätsentwicklung im Bildungswesen. Berlin.

Eckhardt, T. (2019). *Das Bildungswesen in der Bundesrepublik Deutschland 2017/2018 - Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa* (Informationsdossier). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Bonn. https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-dt-pdfs/dossier_de_ebook.pdf

Helmke, A. & Hosenfeld, I. (2003a). Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen. Teil I: Ziele, Konzepte und Organisation. *Schulverwaltung. Ausgabe Hessen, Rheinland-Pfalz und Saarland*, *7*(1), 10–13.

Helmke, A. & Hosenfeld, I. (2003b). Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen. Teil II: Nutzung für Qualitätssicherung und Verbesserung der Unterrichtsqualität. *Schulverwaltung. Ausgabe Hessen, Rheinland-Pfalz und Saarland*, *7*(2), 41–43.

KMK (Hrsg.). (1997). Grundsätzliche Überlegungen zu Leistungsvergleicheninnerhalb der Bundesrepublik Deutschland – Konstanzer Beschluss –. Verfügbar 25. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1997/1997_10_24-Konstanzer-Beschluss.pdf

KMK. (2001, 23. März). *Leistungsmessungen in Schulen.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/leistungsmessungen-in-schulen.html

KMK. (2002a, 1. März). *297. Plenarsitzung der Kultusministerkonferenz am 28. Februar / 01. März 2002 in Berlin.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/297-plenarsitzung-der-kultusministerkonferenz-am-28-februar-01-maerz-2002-in-berlin.html

KMK. (2002b). *298. Plenarsitzung der Kultusministerkonferenz am 23. und 24.Mai 2002 in Eisenach.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/298-plenarsitzung-der-kultusministerkonferenz-am-23-und-24mai-2002-in-eisenach.html

KMK. (2002c, 18. Oktober). *299. Plenarsitzung der Kultusministerkonferenz am 17./18. Oktober 2002 in Würzburg.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/299-plenarsitzung-der-kultusministerkonferenz-am-1718-oktober-2002-in-wuerzburg-1.html

KMK. (2002d). *Nationale Bildungsstandards.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/nationale-bildungsstandards.html

KMK. (2003a, 4. Juli). *Internationale Vergleichsstudie ausgewählter PISA-Teilnehmerstaaten.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/internationale-vergleichsstudie-ausgewaehlter-pisa-teilnehmerstaaten.html

KMK. (2003b). *Qualität in der Bildung braucht die Anstrengung aller.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/qualitaet-in-der-bildung-braucht-die-anstrengung-aller.html

KMK (Hrsg.). (2003d). Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10). Verfügbar 8. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-Vereinbarung-Bildungsstandards-MS.pdf

KMK. (2003e, 4. Juli). *Wolff: Bildungsstandards sind der richtige Weg für mehr Qualität im Unterricht.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/wolff-bildungsstandards-sind-der-richtige-weg-fuer-mehr-qualitaet-im-unterricht.html

KMK. (2006a, 2. Juni). *Ergebnisse der 314.Plenarsitzung der Kultusministerkonferenz.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/ergebnisse-der-314plenarsitzung-der-kultusministerkonferenz.html

KMK (Hrsg.). (2006c). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* LinkLuchterhand.

KMK. (2006d, 18. September). *Länderübergreifendes Kooperationsprojekt 'Lernstandserhebungen in der Grundschule' geht an den Start.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/laenderuebergreifendes-kooperationsprojekt-lernstandserhebungen-in-der-grundschule-geht-an-den-start.html

KMK. (2007, 18. Oktober). *Ergebnisse der 319.Plenarsitzung der Kultusministerkonferenz.* Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/ergebnisse-der-319plenarsitzung-der-kultusministerkonferenz.html

KMK (Hrsg.). (2012b). Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA) (Beschluss der Kultusministerkonferenz vom 08.03.2012). Verfügbar 15. August 2021 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

KMK. (2013). VERA 3 und VERA 8: Fragen und Antworten für Schulen und Lehrkräfte. Verfügbar 15. August 2021 unter https://www.kmk.org/fileadmin/Dateien/pdf/VERA_FragenundAntworten_25-03-2014.pdf

KMK (Hrsg.). (2016b). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Wolters Kluwer.

KMK. (2018a, 15. März). *Qualitätssicherung: Vergleichsarbeiten (VERA) werden modernisiert*. Verfügbar 15. August 2021 unter https://www.kmk.org/presse/pressearchiv/mitteilung/qualitaetssicherung-vergleichsarbeiten-vera-werden-modernisiert.html

KMK (Hrsg.). (2018c). Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA) (Beschluss der Kultusministerkonferenz vom 08.03.2012 i. d. F. vom 15.03.2018). Verfügbar 15. Juli 2020 unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf

Weinert, F. (Hrsg.). (2001b). *Leistungsmessungen in Schulen.* Beltz.

Weinert, F. (2001c). Perspektiven der Schulleistungsmessung - mehrperspektivisch betrachtet. In F. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 353–366). Beltz.

Weinert, F. (2001d). Schulleistungen - Leistungen der Schule oder der Schüler? In F. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 73–86). Beltz.

A.3. Testdomänen, Testheftverteilung und Verbindlichkeit der Vergleichsarbeiten in Berlin

Die folgenden Tabellen geben Auskunft über die in den zurückliegenden Jahren im Rahmen von VERA-3 und VERA-8 überprüften Domänen, über die Verbindlichkeit des Einsatzes und die Verteilung ggf. vorliegender Testheftversionen auf verschiedene Schulformen im Land Berlin.

Tabelle A.2.: Testdomänen und Verbindlichkeit in Berlin bei VERA 3

		Uni Koblenz-Landau				IQB										
Fach	Domäne	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Deutsch	Lesen	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
	Hören								v			v			v	v
	Rechtschreiben					v				v			v			
	Schreiben		v		v		f									
	Sprachgebrauch	v		v				v			v			v		
Mathematik																v
	Größen & Messen	v	v					v	v				v	v		
	Zahlen & Operationen	v	v	v		v	v				v	v				*
	Muster & Strukturen			v			v	v				v	v			
	Raum & Form			v	v				v	v				v	v	*
	Daten, Häufigkeiten ...				v	v				v	v				v	

v = verpflichtete vs. f = freiwillige Teilnahme, vor 2006 wurden die Vergleichsarbeiten in Berlin direkt von der Senatsverwaltung und der Universität Koblenz-Landau betreut; Hierfür liegen keinerlei Daten (mehr) vor. 2006 und 2007 fanden die Vergleichsarbeiten noch in der Klassenstufe 4 statt. 2020 bot das IQB erstmals ein alle Mathematik-Leitideen übergreifenden Mathematiktest an, ergänzte das Erweiterungsmodul mit möglichen Schwerpunkten (*), von denen Berlin nur Zahlen & Operationen nutze.

Tabelle A.3.: Testdomänen, Testheftverteilung und Verbindlichkeit in Berlin bei VERA 8

Fach	Domäne / Heftversion		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Deutsch	Domäne	Lesen		f	f	v	v	v	v	v	v	v	v	v	v
		Hören			f			v				v			v
		Rechtschreiben					kt				v			v	
		Schreiben				kt									
		Sprachgebrauch		f					v				v		
	Version	1	–	○	○	ISS	ISS	○	ISS	ISS	ISS	ISS	ISS	ISS	alle
		2	–	alle	alle	GY	GY	ISS	GY	GY	GY	GY	GY	GY	alle
		3	–	○	○	○	○	GY	○	○	–	–	–	–	–
Mathematik			f	v	v	v	v	v	v	v	v	v	v	v	v
	Version	1	○	○	ISS	ISS	ISS	○	ISS	ISS	ISS	ISS	ISS	ISS	alle
		2	alle	alle	GY	GY	GY	ISS	GY	GY	GY	GY	GY	GY	alle
		3	○	○	○	○	profil	GY	profil	profil	–	–	–	–	–
Englisch	Domäne	Lesen		f	v	v	v	v	v	v	v	v	v	v	v
Französisch		Hören		f		v	v	v	v	v	v	v	v	v	v
		Schreiben			v										
	Version	1	–	○	○	○	ISS	○	ISS	ISS	ISS	ISS	ISS	ISS	alle
		2	–	alle	ISS	ISS	GY	ISS	GY	GY	GY	GY	GY	GY	alle
		3	–	○	GY	GY	bili	GY	bili	bili	–	–	–	–	–

v = verpflichtete vs. f = freiwillige Teilnahme, kt = diese Domäne wurde in Berlin nicht getestet, (–) = dieses Testheft wurde nicht angeboten (2008 wurde nur Mathematik angeboten und ab 2016 wurden für alle Domänen nur noch 2 Testheftversionen angeboten. Ab 2020 konnten aus Module Testhefte zusammengestellt werden, Berlin stellte hier zwei Varianten zusammen.), (○) = Testheft wurde nicht verwendet, profil/bili = Testheft nur in Profil- bzw. bilingualen Klassen verwendet.

A.4. Beschreibung der Testinstrumente für die Domänen Deutsch Lesen und Englisch Leseverstehen der Vergleichsarbeiten in der Jahrgangsstufe 8

Tabelle A.4.: Beschreibung der VERA-8-Testinstrumente für die Domäne Deutsch Lesen

Ver.[a]	Jahr	N	Itemparameter						Testhefte				
			BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung				
			Min	Max	Mw	Sd	Mw	Sd	I	II	III	IV	V
1	2009	28	45	664	433	157	–	–	8	6	8	5	1
	2010	31	228	732	481	110	–	–	6	13	5	3	4
	2011	34	325	769	506	113	0,89	0,63	6	8	11	2	7
	2012	33	242	586	412	92	–	–	12	14	5	2	0
	2013	38	171	627	390	103	–	–	20	10	6	2	0
	2014	40	120	662	384	129	1,11	0,36	24	8	2	4	2
	2015	43	163	755	407	129	1,01	0,32	23	8	8	2	2
2	2009	29	120	664	448	132	–	–	7	8	10	2	2
	2010	31	347	732	517	111	1,30	0,36	4	12	5	3	7
	2011	34	325	780	529	123	0,45	0,47	4	8	11	2	9
	2012	35	248	660	442	117	0,97	0,26	12	10	5	7	1
	2013	35	245	703	453	108	0,90	0,29	11	9	11	3	1
	2014	42	325	669	471	78	0,81	0,43	8	14	14	5	1
	2015	40	147	745	465	133	0,45	0,47	13	12	7	3	5
	2016	36	144	666	407	129	1,12	0,34	21	6	6	2	1
	2017	37	284	722	437	102	0,81	0,45	20	9	4	2	2
	2018	33	165	676	423	128	1,21	0,41	17	8	4	3	1
	2019	41	194	665	447	104	1,12	0,48	17	12	9	2	1
	2020	37	201	695	431	133	–	–	18	7	7	2	3
3	2009	28	126	902	507	148	–	–	5	5	10	4	4
	2010	27	385	727	536	96	–	–	2	6	9	4	6
	2011	36	313	780	547	121	–	–	4	6	12	4	10

Fortsetzung auf der nächste Seite

Tabelle A.4.: Fortsetzung der Tabelle von der Vorseite

Ver.[a]	Jahr	N	Itemparameter						Testhefte				
			BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung				
			Min	Max	Mw	Sd	Mw	Sd	I	II	III	IV	V
	2012	36	263	741	480	105	0,87	0,32	6	12	11	5	2
	2013	40	378	779	533	90	0,80	0,39	1	11	15	8	5
	2014	38	293	723	498	103	–	–	6	11	9	9	3
	2015	46	218	740	513	124	–	–	8	8	15	8	7
	2016	43	315	852	537	138	0,66	0,49	11	8	7	10	7
	2017	35	293	747	518	109	0,74	0,37	7	8	12	6	2
	2018	39	217	829	527	131	0,99	0,48	8	9	10	6	6
	2019	38	246	792	527	142	0,98	0,46	9	5	11	7	6
	2020	34	201	795	518	153	–	–	12	4	7	2	9

[a] Version des Testhefts. Ab dem Jahr 2020 wurden vom IQB nicht genau zwei oder drei Testheftversionen zur Verfügung gestellt, sondern kombinierbare Module. Hier wird jeweils die empfohlene Auswahl des Landes Berlin für die ISS (Testheftversion 2) und die Gymnasien (Testheftversion 3) dargestellt.

[b] Die Diskrimination wurde aus den Daten des regulären Einsatzes bei VERA ermittelt und liegt deshalb für nicht oder nur an kleinen Teilpopulationen eingesetzten Testheftversionen nicht vor.

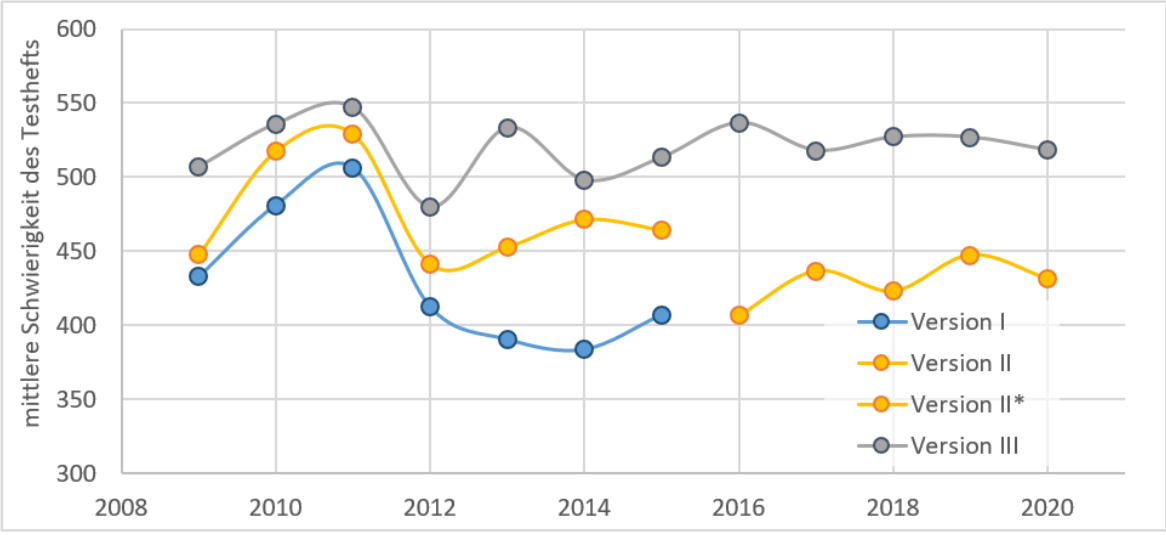


Abbildung A.1.: Mittlere Schwierigkeit der Items für die Testheftversionen bei VERA-8 Deutsch Lesen von 2009 bis 2020

Tabelle A.5.: Beschreibung der VERA-8-Testinstrumente für die Domäne Englisch Leseverstehen

Jahr	N	Itemparameter						Testhefte								
		BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung[c]								
		Min	Max	Mw	Sd	Mw	Sd	A1		A2		B1		B2		C1
Version 1																
2009	61	189	554	380	82	–	–	22	12	15	10	1	1	0	0	0
2010	60	242	557	404	83	–	–	15	8	20	10	6	1	0	0	0
2011	48	92	594	386	127	–	–	16	8	9	7	4	4	0	0	0
2012	42	192	615	412	103	1,45	0,33	9	8	11	9	2	0	3	0	0
2013	56	262	564	428	82	–	–	10	11	11	9	13	2	0	0	0
2014	42	139	552	347	108	1,23	0,39	22	6	7	4	2	1	0	0	0
2015	57	194	615	401	92	1,50	0,46	15	12	14	9	5	0	2	0	0
Version 2[a]																
2009	66	245	607	436	85	1,57	0,55	11	10	20	10	7	7	1	0	0
2010	58	242	649	453	93	1,25	0,43	6	8	16	9	11	5	3	0	0
2011	47	187	626	484	95	1,10	0,52	4	6	6	8	12	7	4	0	0
2012	39	363	649	499	67	1,22	0,34	0	3	6	11	13	2	4	0	0
2013	51	320	632	476	80	1,32	0,64	3	7	8	13	11	6	3	0	0
2014	43	223	573	438	79	1,05	0,44	8	4	11	9	7	4	0	0	0
2015	58	215	661	483	95	1,25	0,47	5	6	10	11	14	3	7	2	0
2016	39	165	853	415	152	1,54	0,39	18	3	1	7	3	3	2	0	2
2017	49	205	669	436	109	1,41	0,52	8	9	13	5	7	3	3	1	0
2018	48	232	589	442	80	1,38	0,50	6	8	9	12	9	4	0	0	0
2019	38	263	609	453	105	1,50	0,35	7	5	5	6	8	4	3	0	0
2020	38	267	632	449	101	–	–	8	7	6	4	5	4	4	0	0
Version 3[a]																
2009	58	286	720	516	91	–	–	1	4	12	9	8	15	6	2	1
2010	52	354	754	544	91	0,82	0,40	0	4	6	8	7	11	12	3	1
2011	44	345	720	558	77	1,06	0,32	1	0	3	4	11	11	11	2	1
2012	41	382	774	564	84	–	–	0	1	1	7	12	7	5	7	1

Fortsetzung auf der nächste Seite

Tabelle A.5.: Fortsetzung der Tabelle von der Vorseite

Jahr	N	Itemparameter						Testhefte								
		BiSta-Werte				Diskrim.[b]		Kompetenzstufenverteilung[c]								
		Min	Max	Mw	Sd	Mw	Sd	A1		A2		B1		B2		C1
2013	43	320	732	564	94	0,96	0,47	2	1	2	5	4	12	11	5	1
2014	42	391	640	516	57	–	–	0	1	6	7	15	11	2	0	0
2015	51	393	819	583	101	1,46	0,57	0	1	4	6	11	5	9	8	7
2016	32	275	853	552	126	1,16	0,34	3	0	1	8	3	6	6	2	3
2017	46	388	801	560	107	1,22	0,47	0	1	9	7	5	7	7	6	4
2018	40	364	760	555	89	1,20	0,55	0	2	3	5	8	11	6	2	3
2019	31	361	846	557	108	1,37	0,30	0	1	4	5	5	5	5	4	2
2020	31	433	715	559	72	–	–	0	0	3	3	7	9	6	1	2

[a] Ab dem Jahr 2020 wurden vom IQB nicht genau zwei Testheftversionen zur Verfügung gestellt, sondern kombinierbare Module. Hier wird jeweils die empfohlene Auswahl des Landes Berlin für die ISS (Testheftversion 2) und die Gymnasien (Testheftversion 3) dargestellt.

[b] Die Diskrimination wurde aus den Daten des regulären Einsatzes bei VERA ermittelt und liegt deshalb für nicht oder nur an kleinen Teilpopulationen eingesetzten Testheftversionen nicht vor.

[c] Die Stufen des GER sind von A1 bis B2 in jeweils zwei Teilstufen differenziert.

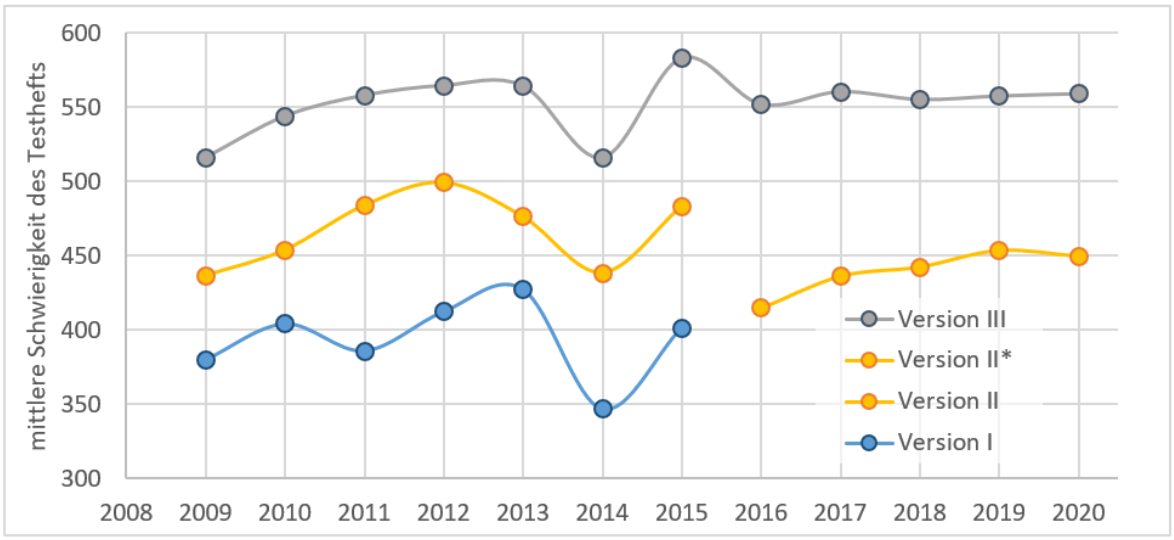
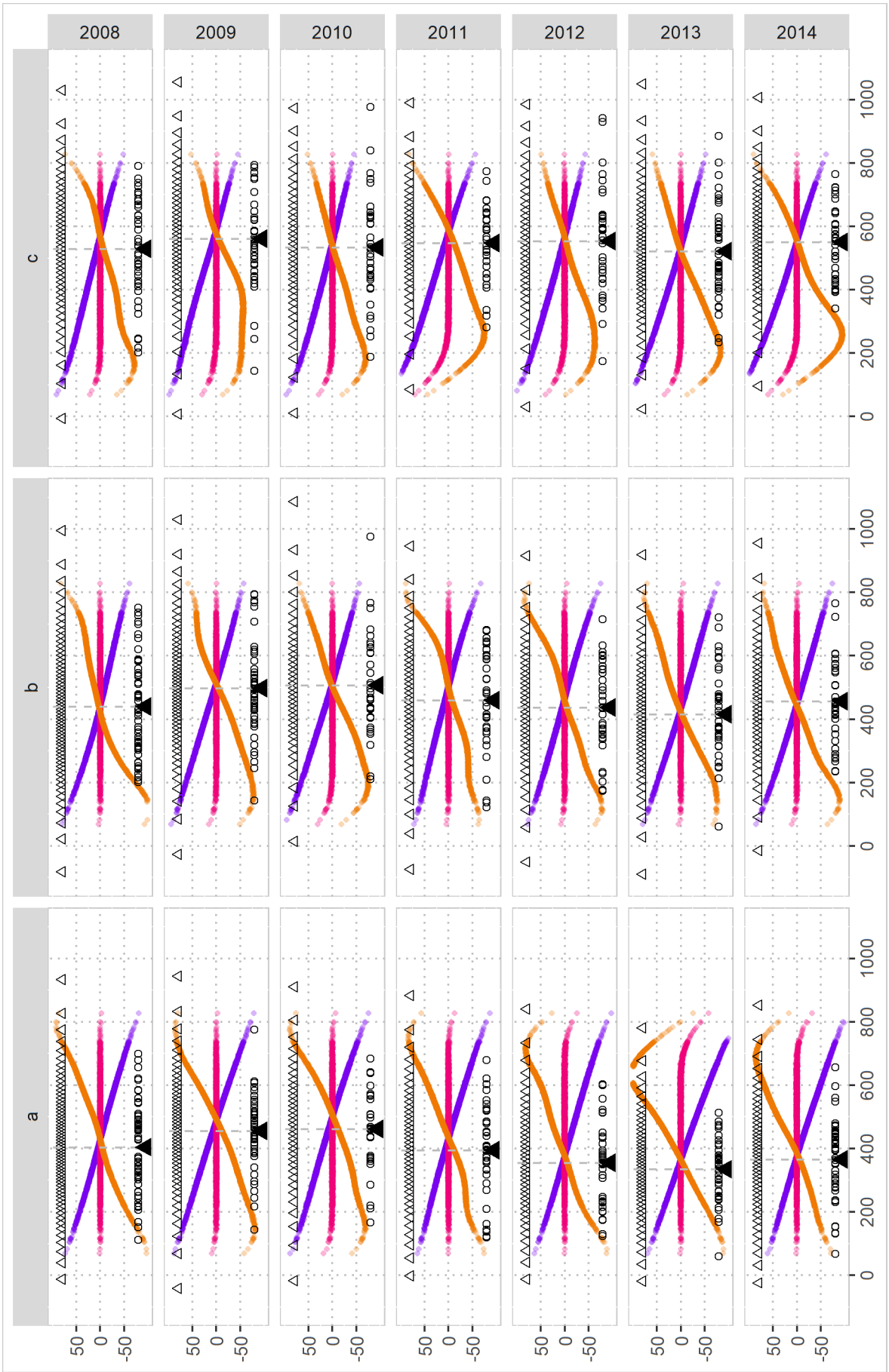


Abbildung A.2.: Mittlere Schwierigkeit der Items für die Testheftversionen bei VERA-8 Englisch Leseverstehen von 2009 bis 2020

A.5. Ergänzende Tabellen und Graphiken zum Kapitel Überprüfung von Gewissheiten beim Einsatz der Rasch-Skalierung

Auf den folgenden Seiten werden zuerst zu den Erläuterungen im Abschnitt 4.4 die Graphiken A.3 und A.4 für das Fach Mathematik ergänzt. Auf die Darstellung der äquivalenten Graphiken für die Domänen Deutsch Lesen sowie Englisch Leseverstehen wurde verzichtet; sie belegen lediglich äquivalente Effekte.

Es folgt dann die vollständige Tabelle zur gekürzten Darstellung als Tabelle 4.6 im Abschnitt 4.4.3.


a
b
c
2008
2009
2010
2011
2012
2013
2014
50
0
-50
0
200
400
600
800
1000

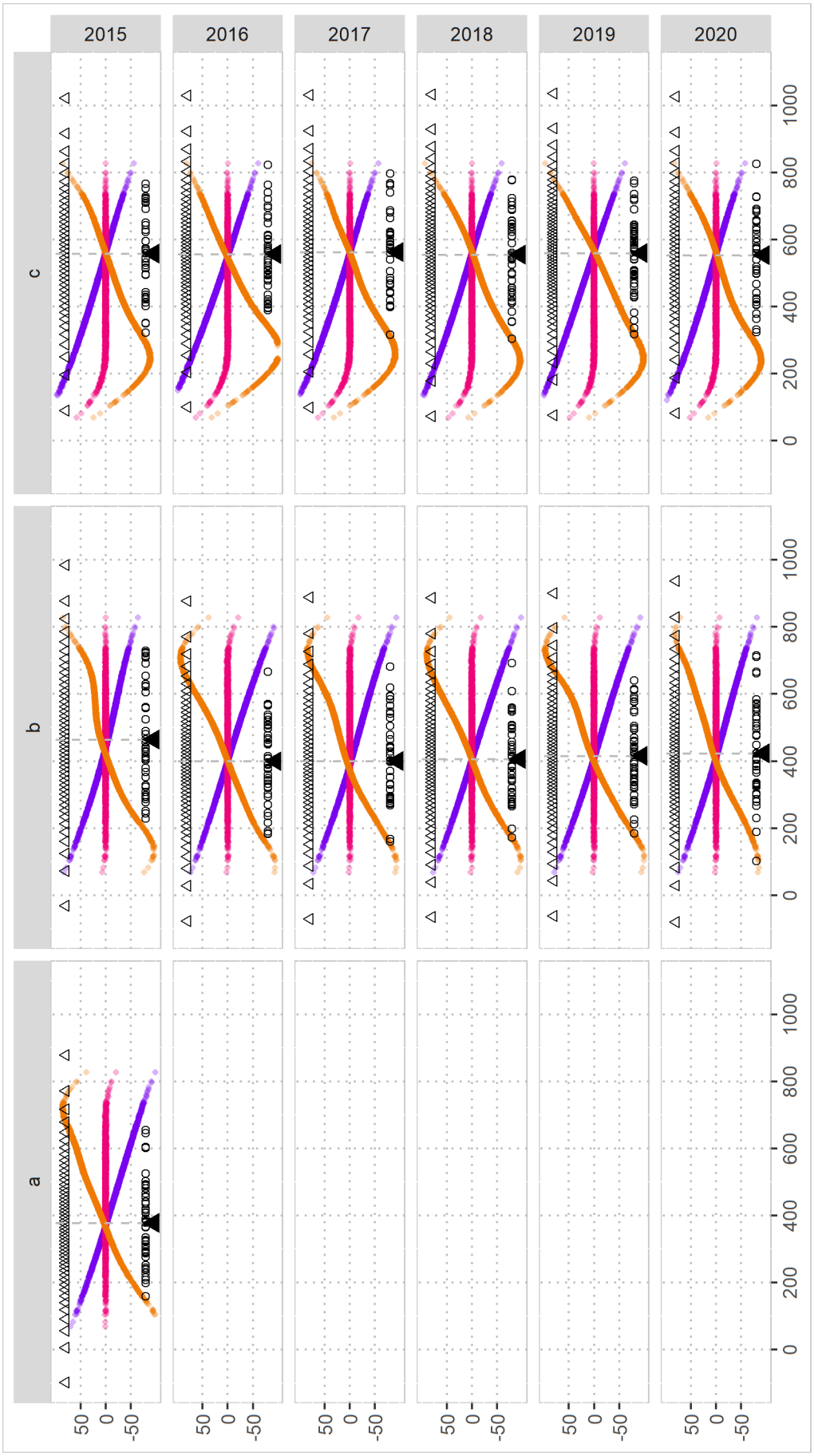
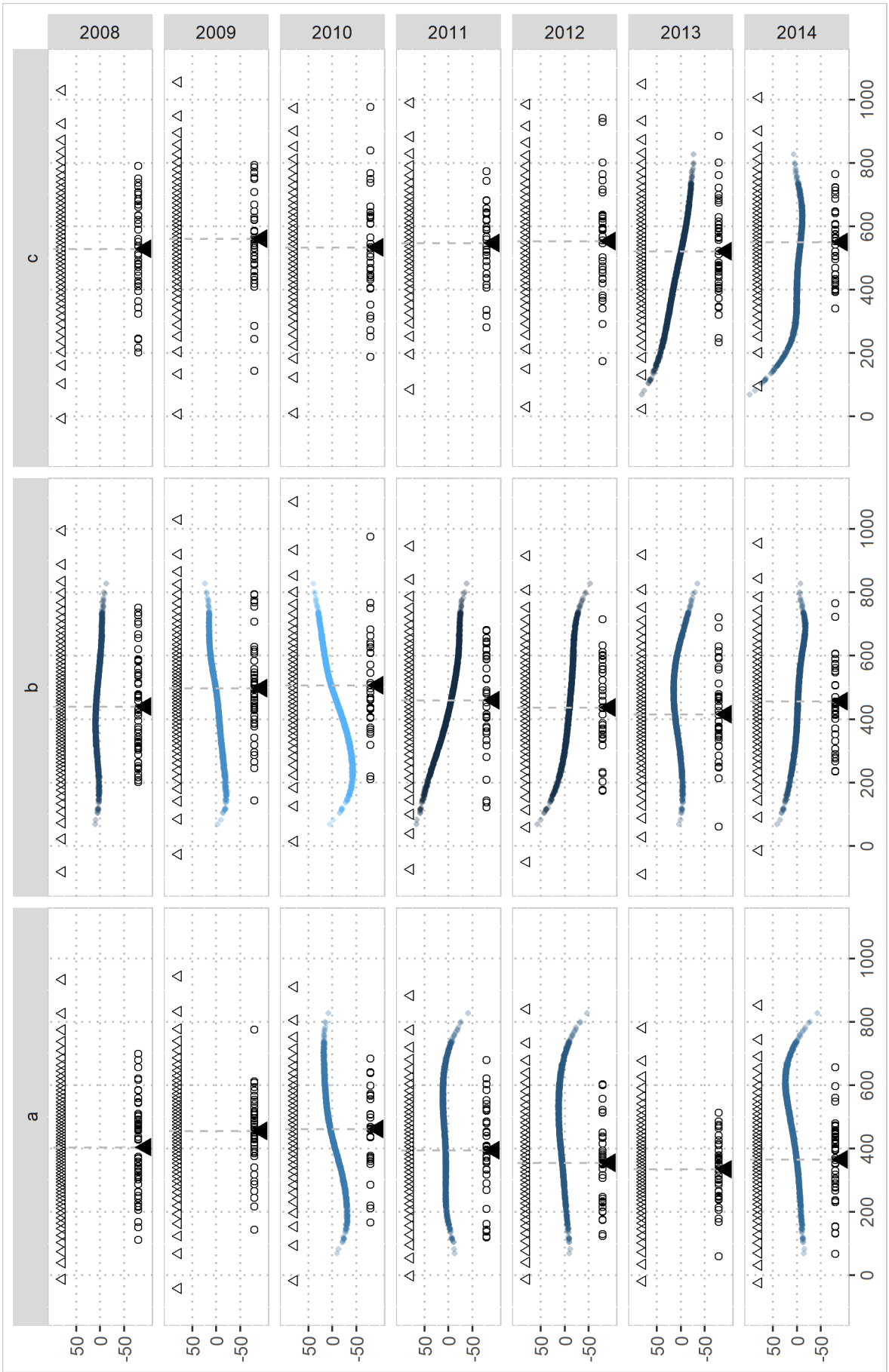


Abbildung A.3.: Differenz von wahrer und geschätzter Fähigkeit für die Testhefte aller Versionen und Jahre für den Fall, dass die Personen die Items mit einer Diskrimination von 0,7 (Orange), 1,0 (Magenta) bzw. 1,7 (Violett) bearbeiten.


a
b
c
2008
2009
2010
2011
2012
2013
2014
50
0
-50
0
200
400
600
800
1000

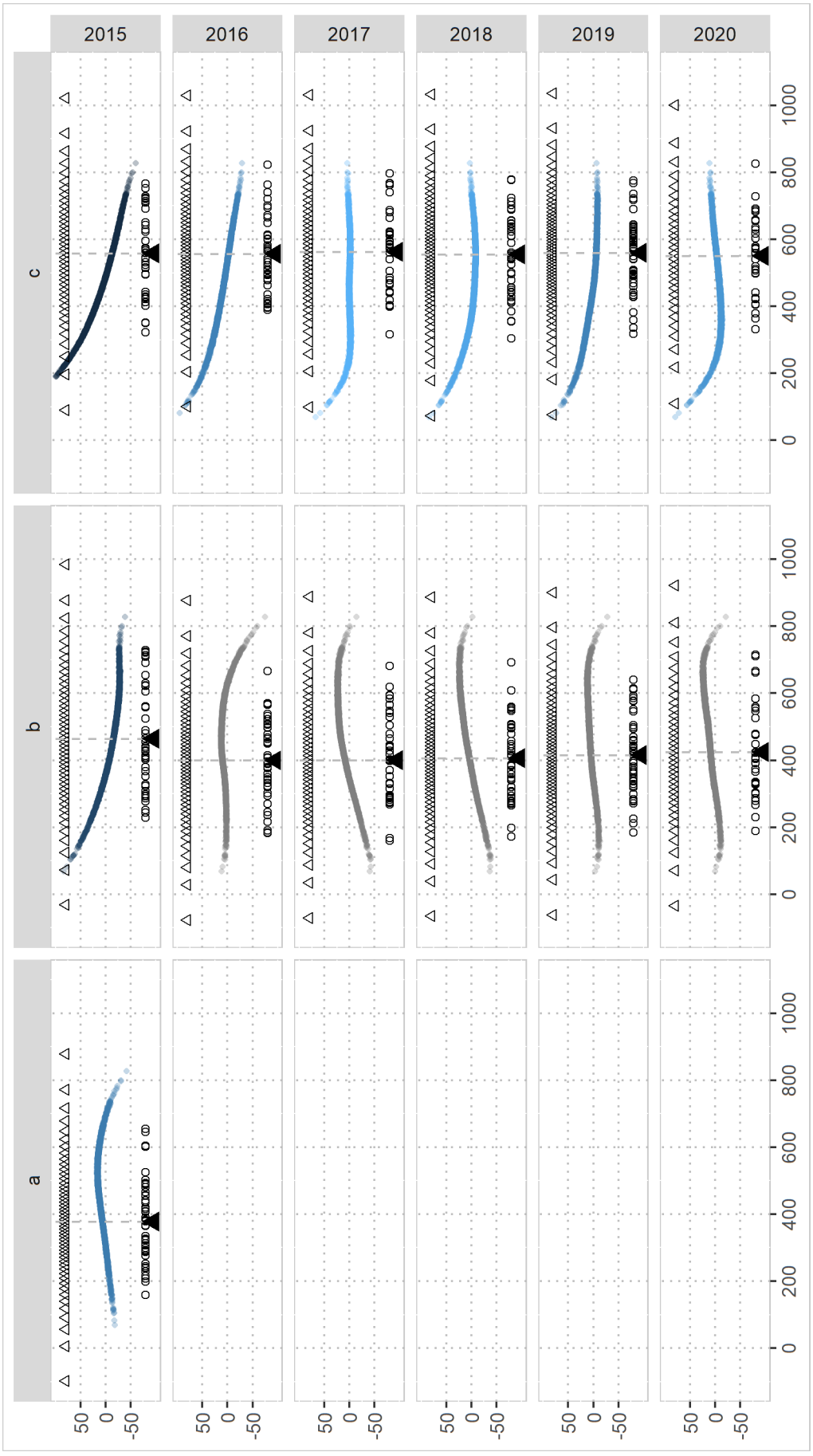


Abbildung A.4.: Differenz von wahrer und geschätzter Fähigkeit für in Berlin tatsächlich eingesetzte Testhefte für den Fall, dass die Personen die Items mit einer Diskrimination bearbeiten, die jener der Berliner Vollerhebung entspricht.

Tabelle A.6.: Anteil der möglichen Personenparameter, die zwischen den zugeordneten Itemparametern liegen, im Bereich ± einer halben und einer Standardabweichung, sowie dem äußeren Bereich der Verteilung der Personenparameter.

				$\pm\frac{1}{2}$ SD			$\pm\frac{1}{2}$ bis 1 SD			außerhalb	
Jahr	Testheft	Items[a]	von[b]	abs.[c]	rel.	von[b]	abs.[c]	rel.	von[b]	abs.[c]	rel.
2008	a	57	22	3	14	18	0	–	16	0	–
2008	b	56	22	8	36	17	0	–	16	0	–
2008	c	42	16	5	31	13	1	8	12	0	–
2009	a	45	18	3	17	14	0	–	12	0	–
2009	b	48	20	2	10	14	0	–	13	0	–
2009	c	40	16	0	–	13	0	–	10	0	–
2010	a	33	13	2	15	11	1	9	8	0	–
2010	b	36	15	1	7	12	0	–	8	1	13
2010	c	36	15	8	53	10	0	–	10	1	10
2011	a	39	16	0	–	12	0	–	10	0	–
2011	b	38	15	7	47	12	0	–	10	1	10
2011	c	32	13	0	–	10	0	–	8	0	–
2012	a	38	15	1	7	12	1	8	10	0	–
2012	b	36	14	3	21	11	1	9	10	0	–
2012	c	34	14	4	29	11	0	–	8	1	13
2013	a	42	17	1	6	14	0	–	10	0	–
2013	b	41	16	0	–	13	0	–	11	0	–
2013	c	42	17	1	6	12	0	–	12	0	–
2014	a	43	18	0	–	12	0	–	12	0	–
2014	b	38	15	0	–	12	0	–	10	0	–
2014	c	38	15	2	13	12	0	–	10	0	–
2015	a	48	19	4	21	15	0	–	13	0	–
2015	b	44	17	1	6	14	0	–	12	0	–
2015	c	38	15	1	7	12	0	–	10	0	–
2016	b	45	18	1	6	14	0	–	12	0	–
2016	c	43	18	1	6	12	0	–	12	0	–
2017	b	41	16	4	25	13	0	–	11	0	–
2017	c	38	15	1	7	12	0	–	10	0	–
2018	b	45	18	1	6	14	0	–	12	0	–
2018	c	47	18	2	11	16	0	–	12	0	–
2019	b	48	19	2	11	15	0	–	13	0	–
2019	c	52	21	1	5	16	0	–	14	0	–
2020	b	46	19	4	21	14	0	–	12	0	–
2020	c	39	16	3	19	12	0	–	10	0	–
Mittelwert		1428	571	77	13%	444	4	1%	379	4	1%

[a] Anzahl der Items. Die Zahl der möglichen Personenparameter ist um einen größer. Weil hier aber die zwei äußeren (approximierten) außen vor bleiben, werden folgend nur Anzahl Items minus 1 Personenparameter einbezogen.
[b] Absolute und relative Anzahl der Personenparameter im entsprechenden Bereich.
[c] Anzahl der Personenparameter, die zwischen den zugeordneten Itemparametern liegen.

A.6. Messzeitpunkte und Ergebnisse des Bildungstrends

Während der ersten Tabelle A.7 zu entnehmen ist, in welchen Jahren welche Domänen im Rahmen des Bildungstrends bzw. im Rahmen der Vergleichsarbeiten überprüft wurden (und werden), geben die Tabelle A.8, A.9 sowie A.10 die Ergebnisse des Bildungstrends wider.

Tabelle A.7.: Messzeitpunkte des Bildungstrends und der Vergleichsarbeiten

			Zyklus 1 - Bildungstrend						Zyklus 2 - Bildungstrend						Zyklus 3 - Bildungstrend			
Fach	Domäne	Test	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Deutsch	Lesen	BT	9		4				9	4					4	9		
		V8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8
		V3		**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**
	Hören	BT	9		4				9	4					4	9		
		V8		8			8			8			8			8		
		V3					3			3			3	3				
	Ortho	BT	9		4				9	4					4	9		
		V8																
		V3		3				3			3							
Mathematik		BT			4	9				4		9			4	9		
		V8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8
		V3		**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**	3	**3**
Fremdspr.	Lesen	BT	9		4				9	4					4	9		
		V8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8
	Hören	BT	9		4				9	4					4	9		
		V8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8	**8**	8

X = bei den fett gedruckten Jahrgangsstufen fand auch eine Pilotierung für das jeweils kommende Jahr statt.
Der Start des 3. Zyklus' war schon für 2020 geplant, wurde aber wegen der Corona-Pandemie um ein Jahr verschoben.

Tabelle A.8.: Kompetenzstufenverteilungen für Deutsch und Mathematik in der Sekundarstufe im Bildungstrend für Berlin

Fach	Domäne	Kohorte	Jahr	Kompetenzstufe					
				Ia[a]	Ib[a]	II	III	IV	V
Deutsch	Lesen	Sek	2009	12,3	16,0	26,5	25,2	14,1	5,9
			2015	13,0	16,3	26,9	25,6	14,4	3,8
			Differenz[b]	0,7	0,3	0,4	0,4	0,3	-2,1
		GY	2009	0,6	3,9	21,1	34,1	27,1	13,1
			2015	1,3	5,7	22,3	35,6	27,1	8,0
			Differenz	0,7	1,8	1,2	1,5	0,0	-5,1
	Hören	Sek	2009	10,5	13,2	20,5	30,7	18,1	7,1
			2015	11,9	13,9	19,3	28	18,1	8,8
			Differenz	1,4	0,7	-1,2	-2,7	0,0	1,7
		GY	2009	0,6	1,8	10,1	37,2	34,4	15,9
			2015	0,8	4,0	12,4	33	31,4	18,3
			Differenz	0,2	2,2	2,3	-4,2	-3,0	2,4
	Orthographie	Sek	2009	6,4	12,6	20,6	33,1	22	5,2
			2015	5,6	12,8	21,9	30,6	21,7	7,4
			Differenz	-0,8	0,2	1,3	-2,5	-0,3	2,2
		GY	2009	0,3	1,1	5,7	37,9	43,3	11,8
			2015	0,2	1,4	8,1	34,0	40,0	16,3
			Differenz	-0,1	0,3	2,4	-3,9	-3,3	4,5
Mathematik		Sek	2011	10,4	22,3	29,9	23,3	11,2	2,8
			2016	9,9	24	27,7	23,2	12,3	2,9
			Differenz	-0,5	1,7	-2,2	-0,1	1,1	0,1
		GY	2011	0,5	6,4	26,1	38,3	22,5	6,2
			2016	0,6	7,6	25,3	36,4	23,6	6,5
			Differenz	0,1	1,2	-0,8	-1,9	1,1	0,3

signifikante Differenzen werden fett dargestellt.

[a]Die Kompetenzstufe I ist bei der Zusammenführung der zwei Kompetenzstufenmodelle für den Hauptschulabschluss (HSA, Jahrgangsstufe 9) und für den Mittleren Schulabschluss (MSA, Jahrgangsstufe 10) zum integrierten Kompetenzstufenmodell in zwei Teilstufen unterteilt worden.

[b]In der Tabelle werden gerundete Werte angegeben. Dadurch kann die Differenz der Anteile minimal von der dargestellten Differenzabweichen (Stanat et al., 2016).

Tabelle A.9.: Kompetenzstufenverteilungen für Englisch in der Sekundarstufe im Bildungstrend für Berlin

Domäne	Kohorte	Jahr	Stufen des gemeinsamen europäischen Referenzrahmens							
			A1.1	A1.2	A2.1	A2.2	B1.1	B1.2	B2.1	≥B2.2[a]
Lesen	Sek	2009	11,1	10,8	14,6	15,9	16,0	14,5	9,9	7,2
		2015	11,6	9,4	11,4	14,4	14,9	14,1	10,8	13,4
		Differenz	0,5	-1,4	**-3,2**	-1,5	-1,1	-0,4	0,9	**6,2**
	GY	2009	0,2	0,6	3,8	11,0	22,7	26,1	20,9	14,7
		2015	0,6	2,2	4,6	10,1	17,5	21,9	18,4	24,6
		Differenz	0,4	1,6	0,8	-0,9	-5,2	-4,2	-2,5	**9,9**
Hören	Sek	2009	3,5	7,7	14,8	19,2	22,1	18,9	9,4	4,3
		2015	3,9	6,7	12,0	16,8	19,3	19,3	12,9	9,1
		Differenz	0,4	-1,0	-2,8	-2,4	-2,8	0,4	3,5	**4,8**
	GY	2009	0,1	0,3	1,5	7,3	26,8	35,6	20,2	8,2
		2015	0,1	0,2	1,7	8,5	18,4	29,5	24,5	17,1
		Differenz	0,0	-0,1	0,2	1,2	**-8,4**	-6,1	4,3	**8,9**

signifikante Differenzen werden fett dargestellt.

[a]Die oberen Stufen B2.2, sowie C1 und C2 werden hier alle zusammengefasst. Da die Tests das Niveau am Ende der Klasse 10 überprüfen, wo in der Regel ein Niveau von B1.2 erwartet wird, differenziert er in den oberen Stufen nur sehr ungenau.

Tabelle A.10.: Kompetenzstufenverteilungen für Deutsch und Mathematik in der Primarstufe im Bildungstrend für Berlin

Fach	Domäne	Jahr	Kompetenzstufe				
			I	II	III	IV	V
Deutsch	Lesen	2011	22,2	24,1	27,3	19,0	7,4
		2016	20,0	23,1	26,4	20,5	10,1
		Differenz	-2,2	-1,0	-0,9	1,5	2,7
	Hören	2011	15,4	23,0	28,5	23,3	9,7
		2016	15,6	23,0	28,7	23,2	9,6
		Differenz	0,2	0,0	0,2	-0,1	-0,1
	Orthographie	2011[a]	-	-	-	-	-
		2016	33,6	27,1	21,5	12,4	5,4
Mathematik		2011	26,6	25,9	23,6	16,5	7,3
		2016	27,6	25,7	23,9	14,4	8,4
		Differenz	1,0	-0,2	0,3	-2,1	1,1

signifikante Differenzen werden fett dargestellt.

[a]Die Stichprobe für die Erfassung von Orthographie im ersten Zyklus war zu klein, so dass auf eine Darstellung im Bericht zum Bildungstrend verzichtet werden musste.

A.7. Abrufquoten der Rückmeldungen

A.7.1. Abrufquoten in Thüringen

Die folgende Tabelle wurde aus den Graphiken der 11 Jahresberichte von 2010 bis 2020 (Nachtigall, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020) erstellt, wobei die Werte dazu aus Graphiken entnommen werden mussten. Durch ein technisches Ablesen auf der Pixelebene konnte eine Genauigkeit von kleiner als +/- 0,2 Prozent erreicht werden. Wiedergegeben werden hier Werte aus zwei Graphiken aus den jeweiligen mit *Rezeption der Testergebnisse an den Schulen* überschriebenen Kapiteln der Jahresberichte, die mit VERA-3 bzw. VERA-8 im Zusammenhang stehen.

Tabelle A.11.: Abrufquoten in Thüringen aus den Jahresberichten von 2010 bis 2020

Jahr	mind. einen Bericht[a]			Bericht mit Vergleich[b]			nach ca. 9 Wo.[c]		nach ca. 18 Wo.[d]		
	GY	GS	RS	GY	GS	RS	DK3	MK3	DK8	MK8	EK8
2008[e]	96,1	96,3	93,6	86,7	84,8	83,3	96,1	95,5	-	92,1	-
2010	97,8	97,6	92,8	79,1	81,3	73,9	78,1	78,0	64,8	68,3	64,9
2011	98,3	97,8	91,6	80,3	77,6	69,7	71,8	73,7	62,1	69,6	64,9
2012	95,1	96,3	90,6	72,4	73,9	65,5	69,8	70,4	56,8	63,0	58,3
2013	97,0	96,5	91,4	71,9	75,4	60,1	71,6	68,8	54,4	58,5	53,9
2014	95,1	92,6	84,3	69,9	71,5	53,9	68,5	64,9	49,8	54,1	48,3
2015	93,9	93,3	82,7	59,4	69,5	51,5	65,5	64,1	46,0	51,0	45,3
2016	92,9	92,8	85,0	56,9	62,8	50,5	59,6	58,5	44,5	51,2	42,8
2017	94,3	91,4	83,2	55,7	65,5	46,5	62,5	60,5	42,8	45,6	41,1
2018	90,4	92,4	78,5	43,8	61,1	42,1	56,5	52,8	34,5	39,7	35,5
2019	93,9	89,9	81,1	49,0	60,4	41,2	55,7	53,5	31,9	39,4	33,6
2020[f]	87,9	-	77,3	40,6	-	37,9	-	-	28,5	36,4	31,4

[a] Anteil an Klassen, die mindestens eine Rückmeldungen abgerufen haben, die es zu verschiedenen Zeitpunkten und für unterschiedliche Fächer gibt. Hierbei sind Rückmeldungen für VERA-6 und -8 einbezogen.

[b] Anteil an Klassen, die in der zweiten Rückmeldewelle mindestens einen Bericht mit korrigierten Landesvergleichsdaten heruntergeladen haben, von denen es pro Fach genau eine gibt. Auch hier sind Rückmeldungen für die Vergleichsarbeiten in der Klassenstufe 6 und 8 subsumiert.

[c] Anteil an Klassen, welche die fachspezifische Rückmeldungen für das Fach Deutsch (DK3) bzw. Mathematik (MK3) für VERA-3 in den ersten 9 Wochen nach Freischaltung heruntergeladen haben.

[d] Anteil an Klassen, welche die fachspezifische Rückmeldungen für das Fach Deutsch (DK8), Mathematik (MK8) oder Englisch (EK8) für VERA-8 in den ersten 18 Wochen nach Freischaltung heruntergeladen haben.

[e] Die Vergleichsarbeiten in der achten Jahrgangsstufe starteten im Jahr 2008 nur mit dem Fach Mathematik.

[f] Auf Grund der Corona-Epedemie fand VERA-3 im gesamten Bundesgebiet nicht statt.

A.7.2. Beispielrückmeldungen VERA-8 Berlin (Englisch)

Die folgenden Bilder repräsentieren eine Beispielrückmeldung für das Fach Englisch des Landes Berlin. Die Rückmeldungen für die anderen Fächer sind äquivalent aufgebaut. Die Bilder zeigen:

- die Sofortrückmeldung (auch klassenbezogene Rückmeldung - Teil 1) mit einer Graphik der Lösungshäufigkeiten aller Einzelaufgaben für jede Domäne[2] und den domänenspezifischen Verteilungen der Schüler*innen auf die Kompetenzstufen (5 Seiten).
- die individuelle Rückmeldung mit einer Gegenüberstellung der Lösungshäufigkeiten für Domänen und Teilkompetenzen des Kindes im Vergleich zu jener der Klasse sowie der klassenbezogenen domänenspezifischen Kompetenzstufenverteilung inkl. der Verortung des Kindes[3] (2 Seiten).
- die klassenbezogene Rückmeldung (Teil 2) mit den Lösungshäufigkeiten für Domänen und Teilkompetenzen für die Klasse, die ganze Schule und eine Vergleichsgruppe und einer schülerweisen Zusammenfassung der Lösungshäufigkeiten und Kompetenzstufenzuordnungen (2 Seiten).
- die Schulrückmeldung mit der Gegenüberstellung der Kompetenzstufenverteilungen aller Klassen und Lerngruppen mit denen der Schule, aller Schüler*innen, die das identische Testheft bearbeitet haben sowie einer Vergleichsgruppe für jede der zwei Domänen sowie einer graphischen Gegenüberstellung der mittleren Lösungshäufigkeiten der Klassen für die Teilkompetenzen jeder Domäne (4 Seiten).

[2] Für das hier dargestellte Fach Englisch sind das die zwei getesteten Domänen Leseverstehen und Hörverstehen, für Deutsch waren es im Jahr 2020 Lesen und Orthographie. Für Mathematik wird nur ein Wert für die globale Mathematikkompetenz zurückgemeldet.

[3] Diese Rückmeldung wird klassenweise erstellt und enthält diese zwei Seiten für jede Schülerin und jeden Schüler.

Bsp. | 8a | Englisch

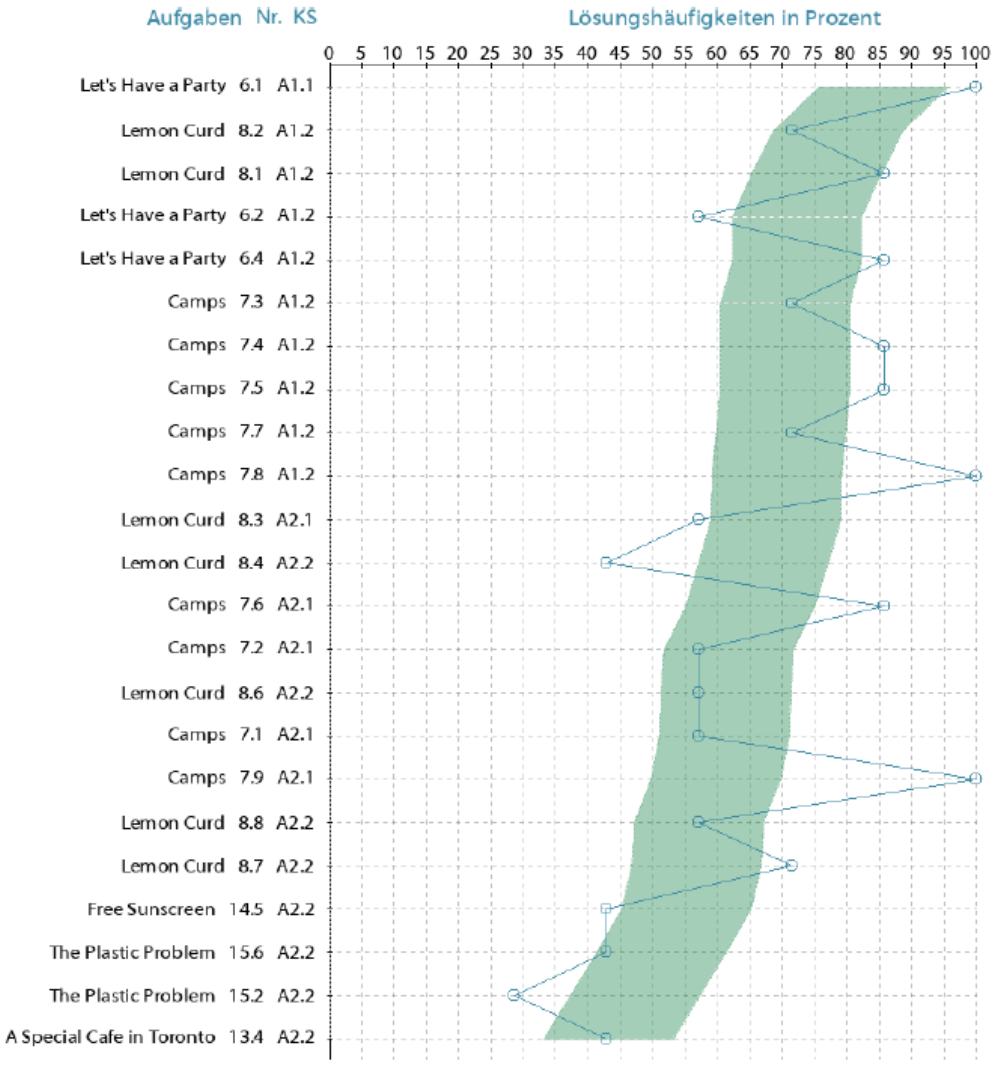
Vera 8 2020

Klassenrückmeldung - Teil 1

Lösungshäufigkeiten einzelner Aufgaben

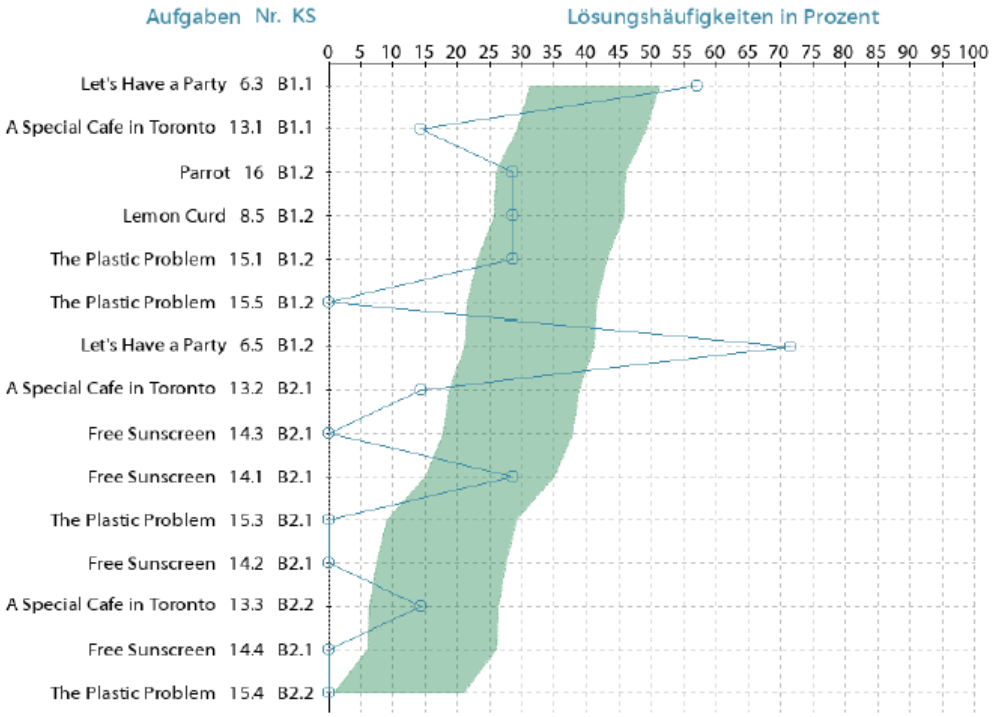
In den folgenden Grafiken sehen Sie die Lösungshäufigkeiten Ihrer Lerngruppe für jede Aufgabe (Linie mit Punkten). Die Aufgaben sind dabei innerhalb der getesteten Kompetenzbereiche aufsteigend nach Ihrer Schwierigkeit sortiert (von der leichtesten bis zur schwierigsten) und zu jeder Aufgabe ist auch die Kompetenzstufe (KS) in Bezug auf die Leistungsniveaus des Gemeinsamen Europäischen Referenzrahmens für Fremdsprachen (GER) angegeben.
Zum Vergleich ist als farbiger Korridor angegeben, wie viel Prozent der Schüler*innen einer länderübergreifenden Stichprobe die entsprechenden Aufgaben durchschnittlich lösen konnten (Pilotierungswerte).

Leseverstehen



Bsp. | 8a | Englisch

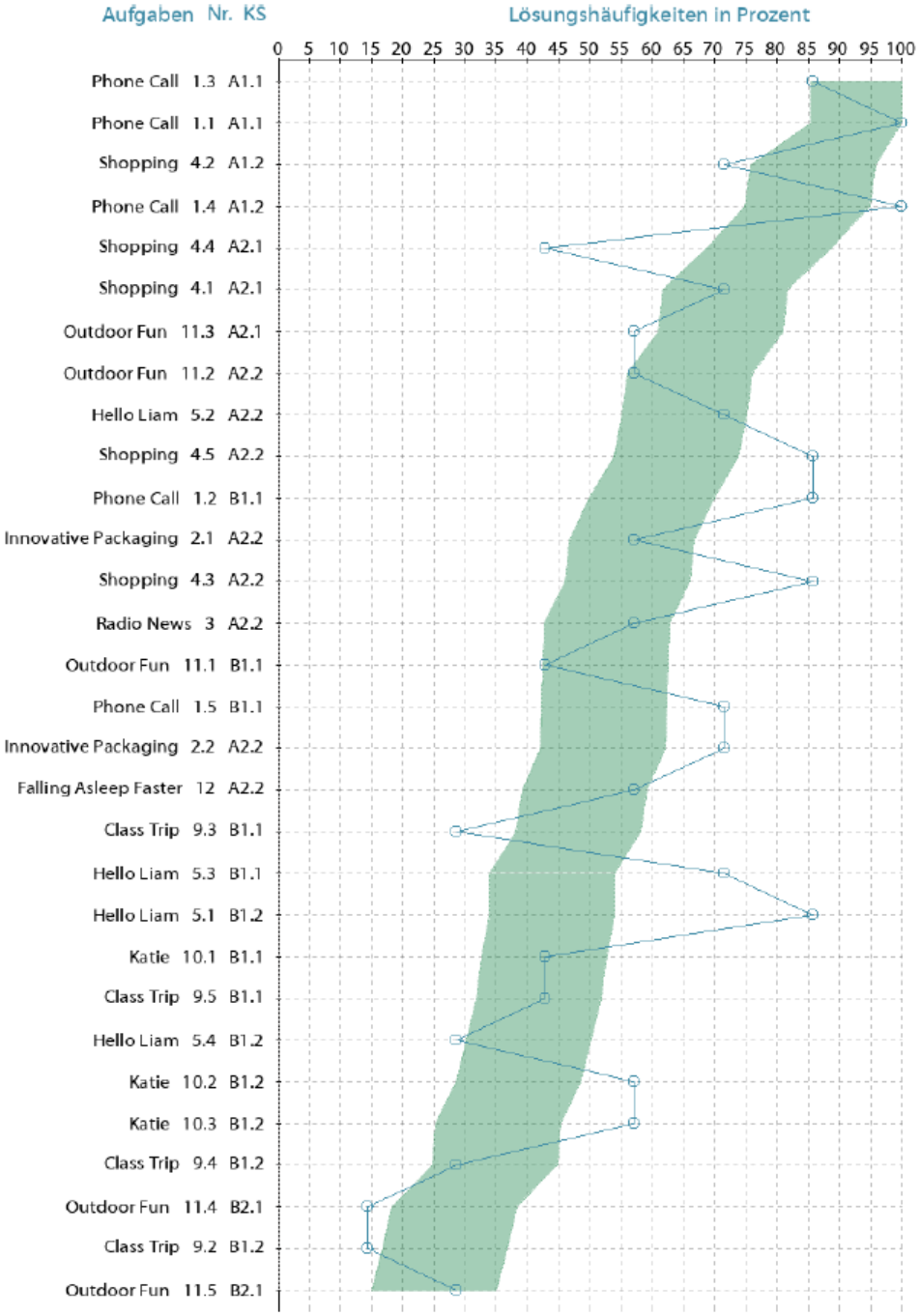
Vera 8 2020



Bsp. | 8a | Englisch

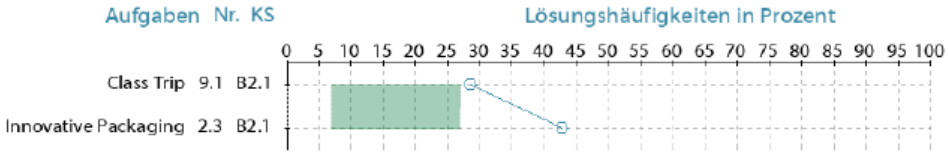
Vera 8 2020

Hörverstehen



Bsp. | 8a | Englisch

Vera 8 2020



Hinweise zur Weiterarbeit

Mit dieser Rückmeldung können Sie besonders auffällige Aufgaben identifizieren, z.B. durch folgende Fragen:

Welche Aufgaben wurden in meiner Lerngruppe häufiger gelöst als in der länderübergreifenden Stichprobe?
Welche Aufgaben wurden in meiner Lerngruppe weniger häufig gelöst als in der länderübergreifenden Stichprobe?
Welche Aufgaben sind inhaltlich auffällig/interessant?

Auffällige Aufgaben können Sie dann, z.B. mithilfe der didaktischen Handreichungen zu VERA 8 analysieren:

Welche Kompetenzen werden mit dieser Aufgabe getestet?
Welche Aufgabenschwierigkeiten bzw. Fehlermuster sind zu erwarten?
Wie kann ich mit dieser Aufgabe im Unterricht weiterarbeiten?

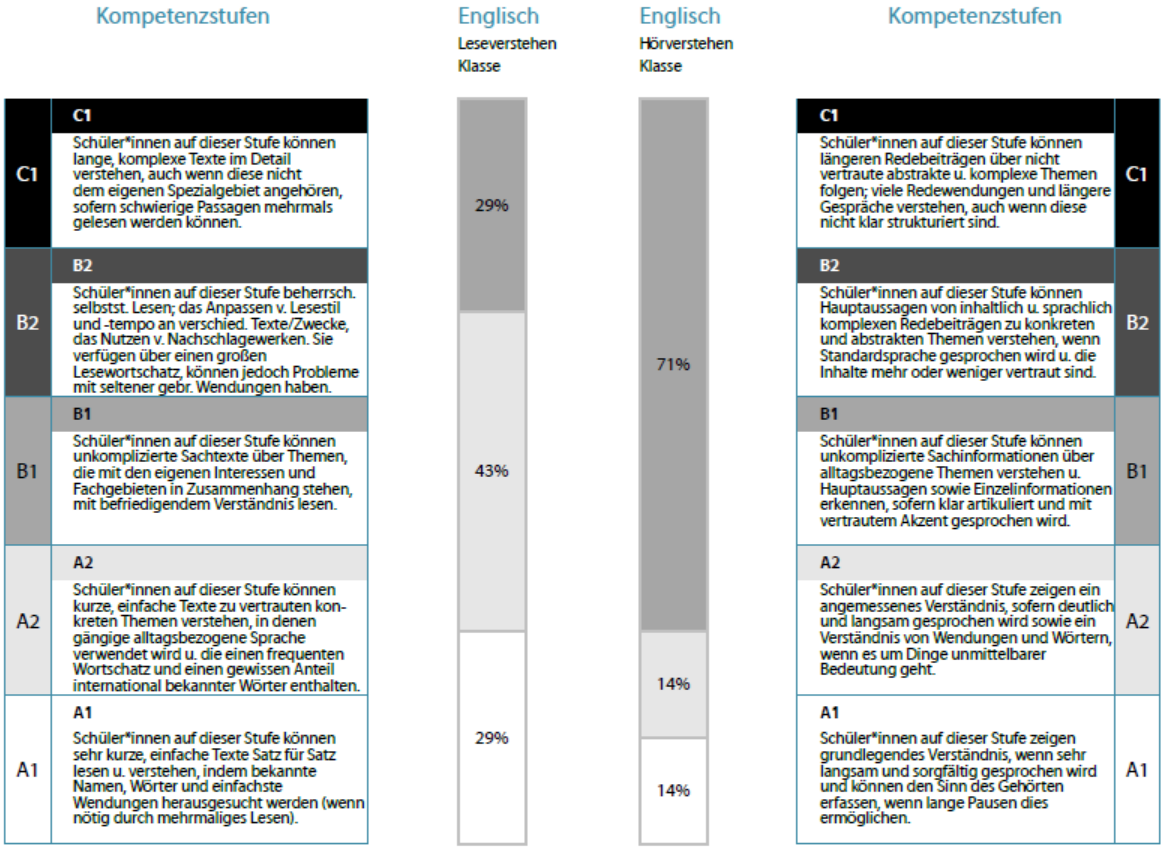
Nutzen Sie dazu gern den ISQ-Aufgabenbrowser, wo Ihnen aktuelle und ältere VERA-Aufgaben inkl. didaktischer Hinweise zur Verfügung gestellt werden (www.aufgabenbrowser.de).

Bsp. | 8a | Englisch

Vera 8 2020

Verteilung der Schüler*innen auf die Kompetenzstufen des GER

Hier erhalten Sie einen Überblick über den Lernstand Ihrer Schülerinnen und Schüler in Bezug auf die Leistungsniveaus des Gemeinsamen Europäischen Referenzrahmens für Fremdsprachen (GER).
In der Grafik sehen Sie neben dem jeweiligen Kompetenzstufenmodell den prozentualen Anteil der Schüler*innen auf den jeweiligen Kompetenzstufen des GER (von A1 - geringster Leistungsstand bis C1 - höchster Leistungsstand).



Mit Bezug auf die bundesweit geltenden Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) gilt:

Kompetenzstufe A1: Die Mindestanforderungen für das Ende der 10. Jahrgangsstufe (MSA) werden nicht erreicht. Zusätzliche zielgerichtete Förderung und Differenzierung ist nötig.
Kompetenzstufe A2: Die Mindestanforderungen für das Ende der 10. Jahrgangsstufe (MSA) werden bereits erreicht und es sind noch zwei Schuljahre Zeit, Kompetenzstufe B1 („Regelstandard") zu erreichen.
Kompetenzstufe B1: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits erreicht („Regelstandard").
Kompetenzstufe B2: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits übertroffen.
Kompetenzstufe C1: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits weit übertroffen.

Schüler*innen auf den Kompetenzstufen B1, B2 und C1 sollten im Unterricht entsprechend Ihres Kompetenzniveaus besonders herausgefordert werden.

Bsp. | 8a | Englisch | Schüler/-in Nr. | .....

Vera 8 2020

Individuelle Rückmeldung

Liebe Eltern,

Ihr Kind hat kürzlich an den Vergleichsarbeiten in der Jahrgangsstufe 8 (VERA 8) teilgenommen und erhält deshalb diese Rückmeldung. Die VERA-8-Ergebnisse werden in erster Linie von den Lehrkräften zur Weiterentwicklung des Unterrichts genutzt. Sie geben Aufschluss darüber, wo die Klasse im Vergleich zu den bundesweit geltenden Bildungsstandards steht, die Schüler*innen am Ende der 10. Jahrgangsstufe erreicht haben sollen. Dadurch kann VERA 8 als eine Art "Frühwarnsystem" gesehen werden.

Bitte bedenken Sie, dass es sich bei jedem Einzelergebnis um eine Momentaufnahme handelt, die von unterschiedlichen Faktoren (z.B. Aufregung am Testtag) beeinflusst worden sein kann. Ein Gesamtbild des Lernstandes Ihres Kindes erhalten Sie nur, wenn Sie auch alle anderen verfügbaren Informationen (z.B. Noten, verbale Beurteilungen) berücksichtigen. Sprechen Sie deshalb mit der Lehrkraft Ihres Kindes, inwieweit die bei VERA 8 erzielten Ergebnisse dem Leistungsstand Ihres Kindes im Unterricht entsprechen.

Wie viele Aufgaben hat Ihr Kind richtig gelöst?

In der folgenden Tabelle* sehen Sie, wie viel Prozent der Aufgaben Ihr Kind richtig gelöst hat. Angegeben ist weiterhin die Lösungshäufigkeit der Lerngruppe Ihres Kindes.

Englisch		Anzahl Schüler/-innen	Anteil richtig gelöster Aufgaben	
			Ihr Kind	Klasse
38 Aufgaben	Leseverstehen (Gesamt)	7	**53%**	**48%**
Lesestile				
	global		**---**	**---**
16	selektiv		**38%**	**35%**
21	detailliert		**67%**	**60%**
	inferierend		**---**	**---**
32 Aufgaben	Hörverstehen (Gesamt)	7	**63%**	**58%**
Hörstile				
	global		**---**	**---**
19	selektiv		**58%**	**53%**
11	detailliert		**64%**	**65%**
	inferierend		**---**	**---**

* In dieser Tabelle werden nur jene Kompetenzen einzeln aufgeführt, welche mit mehr als vier Aufgaben getestet wurden. In das Ergebnis des Gesamttests gehen dagegen alle Aufgaben ein.

ISQ Institut für Schulqualität der Länder Berlin und Brandenburg e.V.

Bsp. | 8a | Englisch | Schüler/-in Nr. | .....

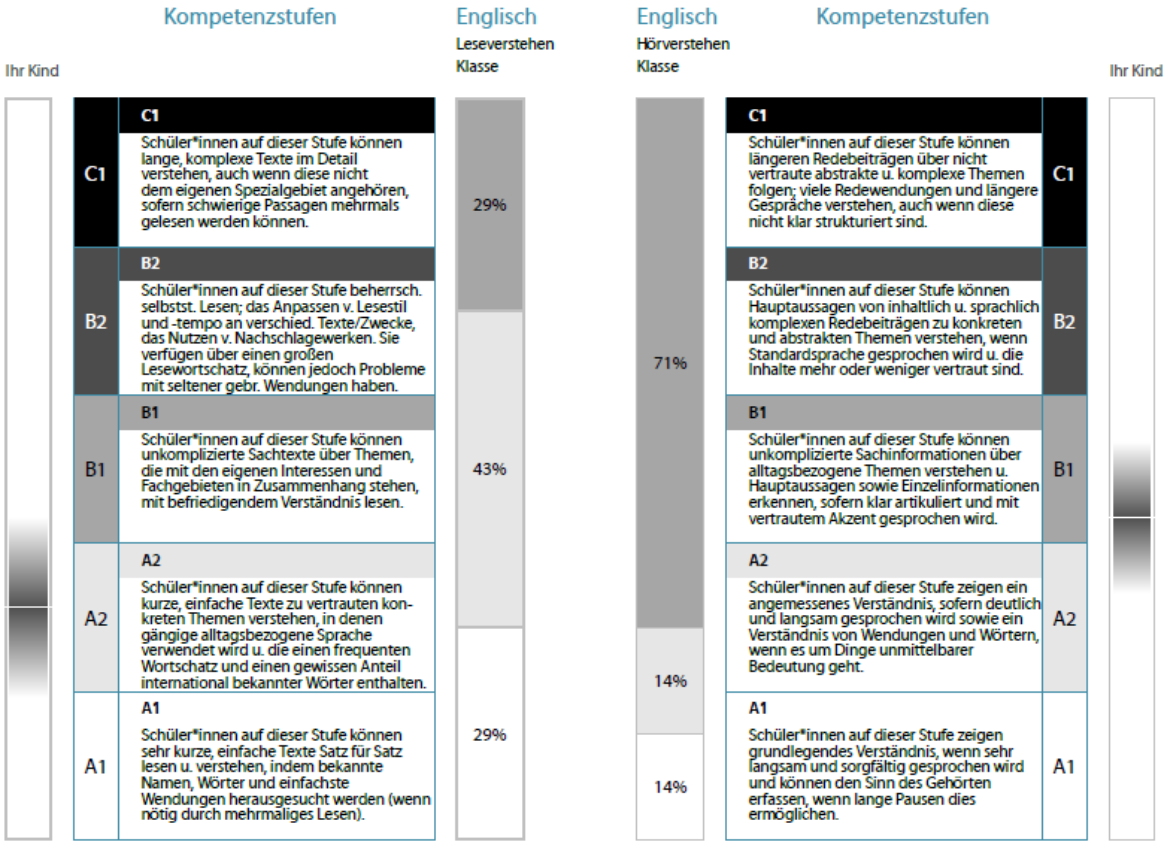
Vera 8 2020

Welche Kompetenzstufe hat Ihr Kind erreicht?

In den bundesweit geltenden "Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10)" sind Stufen ausgewiesen, denen eine Schülerleistung zugeordnet werden kann. Diese Kompetenzstufen beziehen sich auf die im "Gemeinsamen Europäischen Referenzrahmen für Fremdsprachen" (GER) festgelegten Kompetenzniveaus von A1 (geringster Leistungsstand) bis C1 (höchster Leistungsstand). Sie können wie folgt interpretiert werden:

Kompetenzstufe B1, B2, C1: Das ist eine sehr gute Leistung! Die durchschnittlichen Leistungserwartungen für das Ende der 10. Jahrgangsstufe (MSA) wurden bereits erreicht (Stufe B1) oder übertroffen (Stufen B2 & C1).
Kompetenzstufe A2: Das ist eine gute Leistung! Die Mindestanforderungen für das Ende der 10. Jahrgangsstufe (MSA) wurden bereits erreicht und es sind noch zwei Schuljahre Zeit, Stufe B1 zu erreichen.
Kompetenzstufe A1: Hier ist besondere Förderung notwendig, um bis zum Ende der 10. Jahrgangsstufe Stufe B1, also die durchschnittlichen Leistungserwartungen für das Ende der 10. Jahrgangsstufe (MSA) zu erreichen.

In der Grafik sehen Sie, wie Ihr Kind in den beiden getesteten Kompetenzbereichen abgeschnitten hat (Balken "Ihr Kind"). Die Zuordnung zu einer Kompetenzstufe ist dabei nur mit einer gewissen Unschärfe möglich, die hier als Farbverlauf dargestellt ist. Angegeben ist weiterhin, wie viel Prozent der Schüler*innen in der Lerngruppe Ihres Kindes die einzelnen Kompetenzstufen erreicht haben (Balken "Klasse").



Bsp. | 8a | Englisch

Vera 8 2020

Klassenrückmeldung (Teil 2) mit Vergleichswerten

Lösungshäufigkeiten für einzelne Kompetenzen

In der Tabelle* sehen Sie für die getesteten Kompetenzen, wie viel Prozent der Aufgaben in Ihrer Lerngruppe richtig gelöst wurden (Spalte "Ihre Klasse").
Zum Vergleich ist außerdem angegeben, ...
- wie viel Prozent der Aufgaben an Ihrer Schule richtig gelöst wurden (Spalte "Schule") und
- wie viel Prozent der Aufgaben von allen Berliner Schüler*innen, die das gleiche Testheft bearbeitet haben wie Ihre Lerngruppe, richtig gelöst wurden (Spalte „Berlin - gleiches TH“).

Englisch		Anteil richtig gelöster Aufgaben								
		Ihre Klasse			Schule			Berlin - gleiches TH		
Aufgaben		alle	männl.	weibl.	alle	männl.	weibl.	alle	männl.	weibl.
	Anzahl Schüler/innen	(7)	(4)	(3)	(76)	(42)	(34)	(10429)	(5578)	(4849)
38	Leseverstehen (Gesamt)	**48%**	45%	53%	**52%**	51%	54%	**51%**	48%	53%
Lesestile										
	global	---	---	---	---	---	---	---	---	---
	selektiv	**35%**	31%	40%	**38%**	38%	39%	**39%**	37%	41%
	detailliert	**60%**	58%	62%	**64%**	62%	66%	**61%**	58%	64%
	Anzahl Schüler/innen	(7)	(4)	(3)	(76)	(42)	(34)	(10417)	(5571)	(4844)
32	Hörverstehen (Gesamt)	**58%**	59%	55%	**57%**	58%	55%	**54%**	53%	55%
Hörstile										
	global	---	---	---	---	---	---	---	---	---
	selektiv	**53%**	54%	53%	**55%**	56%	53%	**52%**	51%	54%
	detailliert	**65%**	68%	61%	**57%**	59%	55%	**55%**	53%	57%

* In dieser Tabelle werden nur jene Kompetenzen einzeln aufgeführt, welche mit mehr als vier Aufgaben getestet wurden. In das Ergebnis des Gesamttests gehen dagegen alle Aufgaben ein.

Die Ergebnisse geben Ihnen Hinweise auf Stärken und Schwächen Ihrer Lerngruppe und damit auf mögliche Schwerpunktsetzungen zur weiteren Förderung. Als Unterstützungsangebote stehen Ihnen das Selbstevaluationsportal (SEP) (Einschätzung der Unterrichtsqualität unter www.sep.isq-bb.de) und der Aufgabenbrowser (aktuelle und ältere VERA Aufgaben inkl. didaktischer Kommentierungen unter www.aufgabenbrowser.de) zur Verfügung.

ISQ Institut für Schulqualität der Länder Berlin und Brandenburg e.V.

Bsp. | 8a | Englisch

Vera 8 2020

Ergebnisse für einzelne Schüler*innen

In dieser Tabelle* sehen Sie für die getesteten Kompetenzen, wie viel Prozent der Aufgaben jede einzelne Schülerin/jeder einzelne Schüler richtig gelöst hat und welche Kompetenzstufe erreicht wurde.
Anhand der Schülerlisten, die an Ihrer Schule vorliegen, können Sie einzelnen Schülerinnen und Schülern ihre Testergebnisse zuordnen. Sie erhalten damit einen differenzierten Überblick über die Einzelleistungen in Ihrer Lerngruppe.

* In dieser Tabelle werden nur jene Kompetenzen einzeln aufgeführt, welche mit mehr als vier Aufgaben getestet wurden. In das Ergebnis des Gesamttests gehen dagegen alle Aufgaben ein.

	Name	Geschlecht	Verkehrssprache	teilnahmeverpflichtet (t)	Sonderpäd. Förderschwerpunkt (SFSP)	Testheftversion	Leseverstehen					Hörverstehen				
							Lesestile					Hörstile				
							global in %	selektiv in %	detailliert in %	Leseverstehen Gesamt in %	Kompetenzstufe	global in %	selektiv in %	detailliert in %	Hörverstehen Gesamt in %	Kompetenzstufe
	8a (Gesamt)						---	**35**	**60**	**48**	---	---	**53**	**65**	**58**	---
1		m	a	x	kein	rTH	---	25	57	42	A2	---	53	82	63	B1
2		m	a		kein	rTH	---	---	---	---	---	---	---	---	---	---
3		m	a	x	kein	rTH	---	38	67	53	A2	---	58	64	63	B1
4		m	a	x	kein	rTH	---	---	---	---	---	---	---	---	---	---
5		w	a	x	kein	rTH	---	---	---	---	---	---	---	---	---	---
6		m	D	x	kein	rTH	---	---	---	---	---	---	---	---	---	---
7		w	D	x	kein	rTH	---	63	76	71	B1	---	79	82	81	B1
8		m	D	x	kein	rTH	---	---	---	---	---	---	---	---	---	---
9		w	D	x	kein	rTH	---	44	71	61	A2	---	58	82	66	B1
10		w	D	x	kein	rTH	---	13	38	26	A1	---	21	18	19	A1
11		m	D	x	kein	rTH	---	44	86	66	B1	---	74	73	75	B1
12		m	D	x	kein	rTH	---	19	24	21	A1	---	32	55	38	A2

Bsp | Schulrückmeldung | Englisch

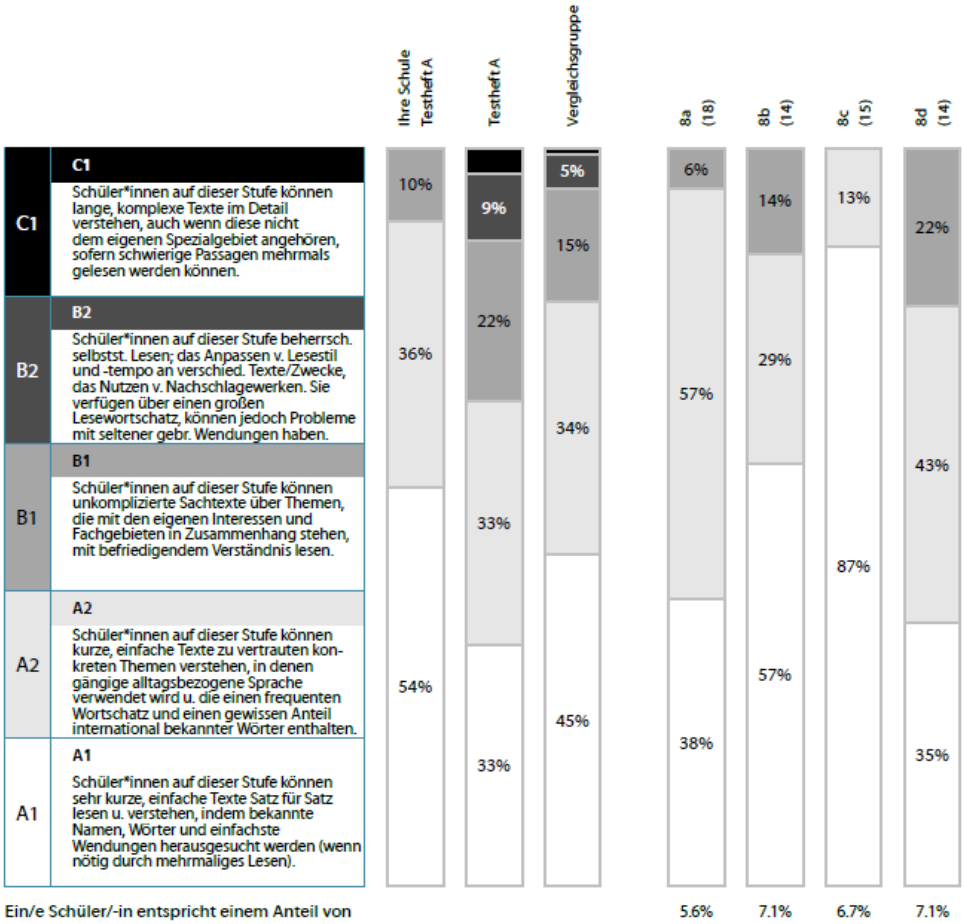
Vera 8 2020

Schulrückmeldung

Hier sehen Sie verschiedene Kompetenzstufenverteilungen nebeneinander. Es werden die Kompetenzstufenverteilung der gesamten Schülerschaft Ihrer Schule (Balken „Ihre Schule") sowie der einzelnen Lerngruppen (Balken mit Klassen/ Kursnamen) dargestellt.
Bitte beachten Sie bei der Interpretation der Kompetenzstufenverteilungen, auf wie viele Schüler*innen sich die prozentualen Angaben jeweils beziehen. Sie können jeweils unter den Balken ablesen, wie viel Prozent in etwa einer Schülerin/einem Schüler entsprechen.

Zur besseren Einordnung der Ergebnisse sind zwei **Vergleichshorizonte** angegeben
1. die Kompetenzstufenverteilung für alle Berliner Schüler*innen, die das gleiche Testheft bearbeiteten wie die Schüler*innen an Ihrer Schule (Balken „Testheft") und
2. die Kompetenzstufenverteilung einer Vergleichsgruppe aus sechs Schulen, die Ihrer Schule in spezifischen Rahmenbedingungen (Anteil der Schüler*innen mit Lernmittelbefreiung und nichtdeutscher Herkunftssprache) sehr ähnlich sind (Balken „Vergleichsgruppe").

Kompetenzstufen - Leseverstehen

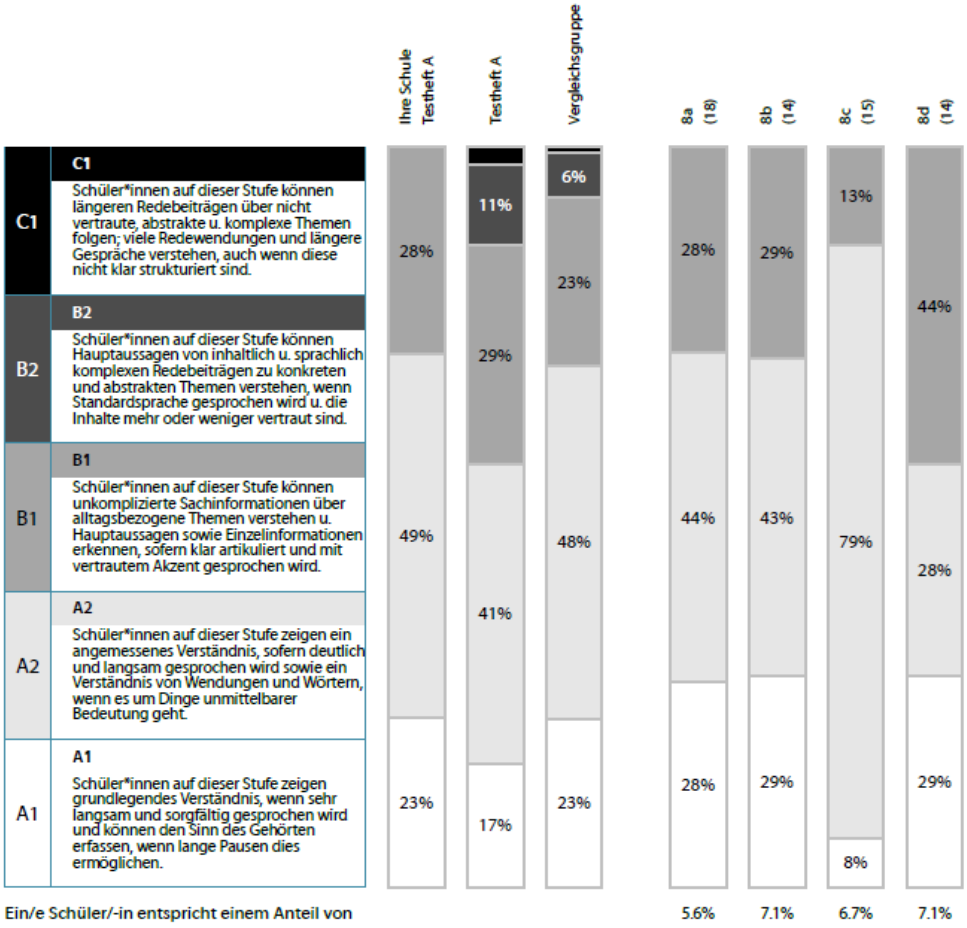


ISQ Institut für Schulqualität der Länder Berlin und Brandenburg e.V.

Bsp | Schulrückmeldung | Englisch

Vera 8 2020

Kompetenzstufen - Hörverstehen



Zur Bedeutung der Kompetenzstufen

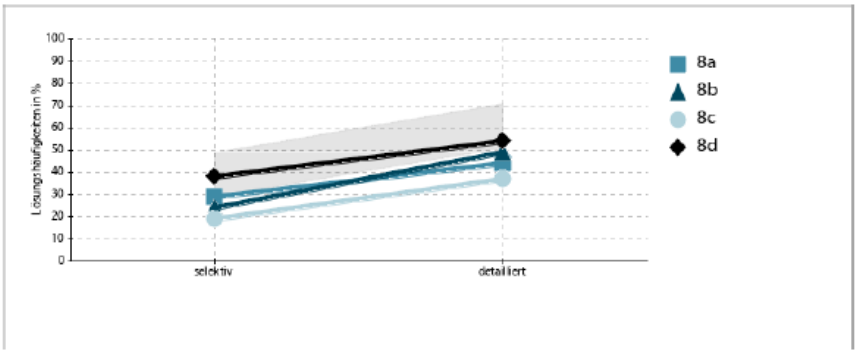
In VERA 8 wird das Erreichen der „Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10)" schon in der 8. Jahrgangsstufe überprüft, um einen ausreichenden Handlungsspielraum für differenzierte, zielgerichtete Interventionen im Unterricht zu eröffnen. Mit Bezug auf die bundesweit geltenden Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) gilt:
Kompetenzstufe A1: Die Mindestanforderungen für das Ende der 10. Jahrgangsstufe (MSA) werden nicht erreicht. Zusätzliche zielgerichtete Förderung und Differenzierung ist nötig.
Kompetenzstufe A2: Die Mindestanforderungen für das Ende der 10. Jahrgangsstufe (MSA) werden bereits erreicht und es sind noch zwei Schuljahre Zeit, Kompetenzstufe B1 („Regelstandard") zu erreichen.
Kompetenzstufe B1: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits erreicht („Regelstandard").
Kompetenzstufe B2: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits übertroffen.
Kompetenzstufe C1: Die durchschnittlichen Leistungserwartungen der Bildungsstandards für das Ende der 10. Jahrgangsstufe (MSA) werden bereits weit übertroffen.

Schüler*innen auf den Kompetenzstufen B1, B2 und C1 sollten im Unterricht entsprechend Ihres Kompetenzniveaus besonders herausgefordert werden.

Lösungshäufigkeiten für Kompetenzen: Lerngruppen Ihrer Schule im Vergleich

Im Folgenden sehen Sie für die getesteten Kompetenzen, wie viel Prozent der Aufgaben in den einzelnen Lerngruppen jeweils richtig gelöst wurden*. Somit können Sie insbesondere Unterschiede zwischen den Lerngruppen auf einen Blick erkennen.
Zur besseren Einordnung der Ergebnisse ist erneut ein **Vergleichshorizont** angegeben: Der grau hinterlegte Bereich gibt eine Spannweite der durchschnittlichen Lösungshäufigkeiten aller Berliner Schüler*innen an, die das gleiche Testheft bearbeiteten wie die Schüler*innen an Ihrer Schule. Liegen die Werte einer Lerngruppe außerhalb dieses Bereiches, handelt es sich um bedeutsame Abweichungen von diesem Durchschnittswert.

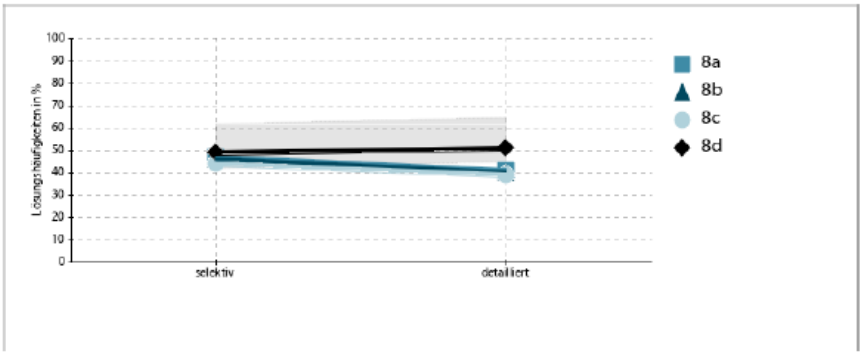
Leseverstehen - Testheft A



Bsp | Schulrückmeldung | Englisch

Vera 8 2020

Hörverstehen - Testheft A



*Hier werden nur jene Kompetenzen einzeln aufgeführt, welche mit mehr als vier Aufgaben getestet wurden. In das Ergebnis des Gesamttests gehen dagegen alle Aufgaben ein.

A.7.3. Unterschiede zwischen den Schulformen

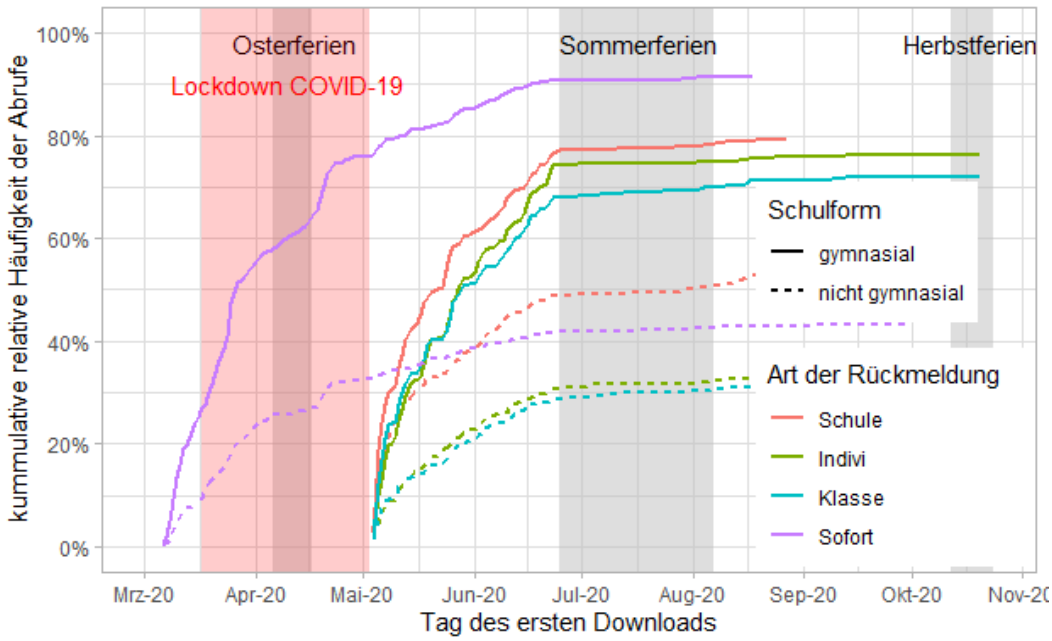


Abbildung A.5.: Downloads von Rückmeldungen für das Fach Deutsch, nach Schulform

Tabelle A.12.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Deutsch

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-0,2678	0,0582	-4,6012	0,0000	0,7651	[0,6824; 0,8573]
	Schulform	2,6512	0,1502	17,6555	0,0000	14,1717	[10,6459; 19,1978]
	R^2	Cohens d			Deviance	dof	AIC
	0,2256	0,5398			2035,78	1875	2039,78
Indivi	(Intercept)	-0,6399	0,0607	-10,5488	0,0000	0,5273	[0,4678; 0,5935]
	Schulform	1,8089	0,1090	16,6021	0,0000	6,1040	[4,9403; 7,5740]
	R^2	Cohens d			Deviance	dof	AIC
	0,1537	0,4262			2288,60	1875	2292,60
Klasse	(Intercept)	-0,7448	0,0617	-12,0643	0,0000	0,4748	[0,4204; 0,5355]
	Schulform	1,6966	0,1058	16,0431	0,0000	5,4553	[4,4412; 6,7235]
	R^2	Cohens d			Deviance	dof	AIC
	0,1405	0,4042			2309,12	1875	2313,12
Schule	(Intercept)	0,1652	0,1201	1,3754	0,1690	1,1797	[0,9326; 1,4945]
	Schulform	1,1912	0,2268	5,2530	0,0000	3,2910	[2,1290; 5,1891]
	R^2	Cohens d			Deviance	dof	AIC
	0,0903	0,3151			553,21	443	557,21

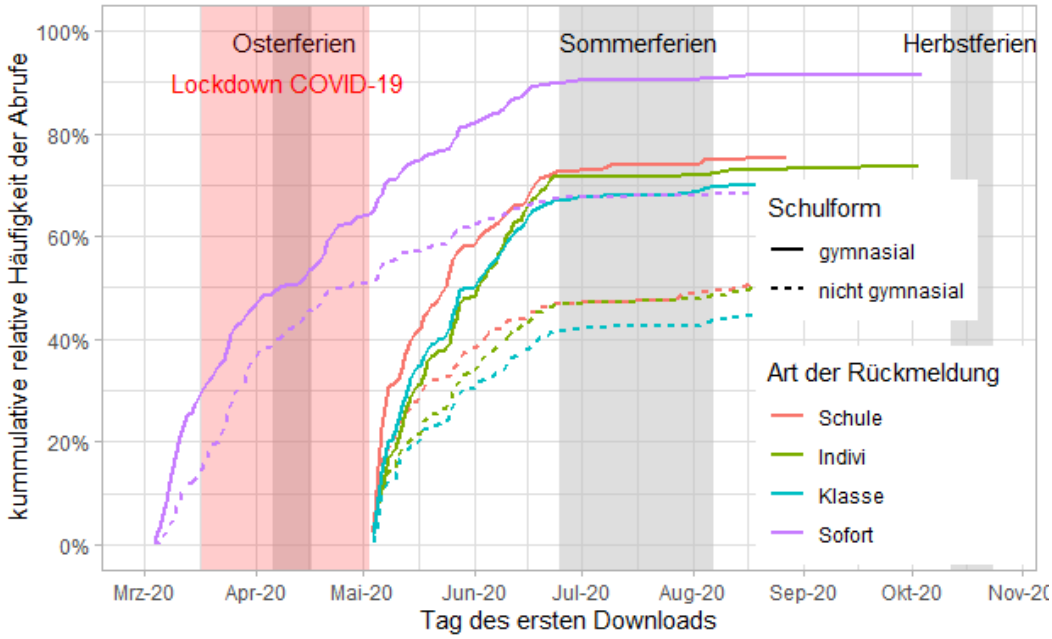


Abbildung A.6.: Downloads von Rückmeldungen für das Fach Englisch, nach Schulform

Tabelle A.13.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Englisch

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	0,8337	0,0614	13,5874	0,0000	2,3018	[2,0429; 2,5986]
	Schulform	1,5478	0,1526	10,1462	0,0000	4,7012	[3,5144; 6,3976]
	R^2	Cohens d			Deviance	dof	AIC
	0,0669	0,2677			1926,74	1918	1930,74
Indivi	(Intercept)	0,0986	0,0565	1,7473	0,0806	1,1037	[0,9881; 1,2330]
	Schulform	0,9404	0,1049	8,9614	0,0000	2,5611	[2,0883; 3,1515]
	R^2	Cohens d			Deviance	dof	AIC
	0,0435	0,2132			2501,46	1918	2505,46
Klasse	(Intercept)	-0,1401	0,0565	-2,4791	0,0132	0,8692	[0,7779; 0,9710]
	Schulform	1,0280	0,1025	10,0285	0,0000	2,7956	[2,2898; 3,4228]
	R^2	Cohens d			Deviance	dof	AIC
	0,0539	0,2387			2536,87	1918	2540,87
Schule	(Intercept)	0,0946	0,1207	0,7836	0,4333	1,0992	[0,8677; 1,3937]
	Schulform	1,0368	0,2183	4,7500	0,0000	2,8201	[1,8509; 4,3614]
	R^2	Cohens d			Deviance	dof	AIC
	0,0726	0,2797			562,83	437	566,83

A.7.4. Unterschiede zwischen den Ländern

Tabelle A.14.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Mathematik für Berlin und Brandenburg

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-0.4125	0.0665	-6.2050	0.0000	0.6620	0.5807; 0.7537
	Land	1.9166	0.1045	18.3399	0.0000	6.7979	5.5485; 8.3588
	R^2	Cohens d			Deviance	dof	AIC
	0.1754	0.4612			2249.86	1976	2253.86
Indivi	(Intercept)	-0.7983	0.0703	-11.3481	0.0000	0.4501	[0.3916; 0.5160]
	Land	1.5829	0.0972	16.2879	0.0000	4.8689	[4.0289; 5.8974]
	R^2	Cohens d			Deviance	dof	AIC
	0.1348	0.3948			2455.22	1976	2459.22
Klasse	(Intercept)	-1.0156	0.0737	-13.7858	0.0000	0.3622	[0.3130; 0.4178]
	Land	1.5502	0.0979	15.8393	0.0000	4.7124	[3.8945; 5.7161]
	R^2	Cohens d			Deviance	dof	AIC
	0.1283	0.3837			2455.49	1976	2459.49
Schule	(Intercept)	0.0183	0.1355	0.1355	0.8923	1.0185	[0.7808; 1.3289]
	Land	0.6161	0.1942	3.1725	0.0015	1.8518	[1.2676; 2.7156]
	R^2	Cohens d			Deviance	dof	AIC
	0.0304	0.1770			596.43	444	600.43

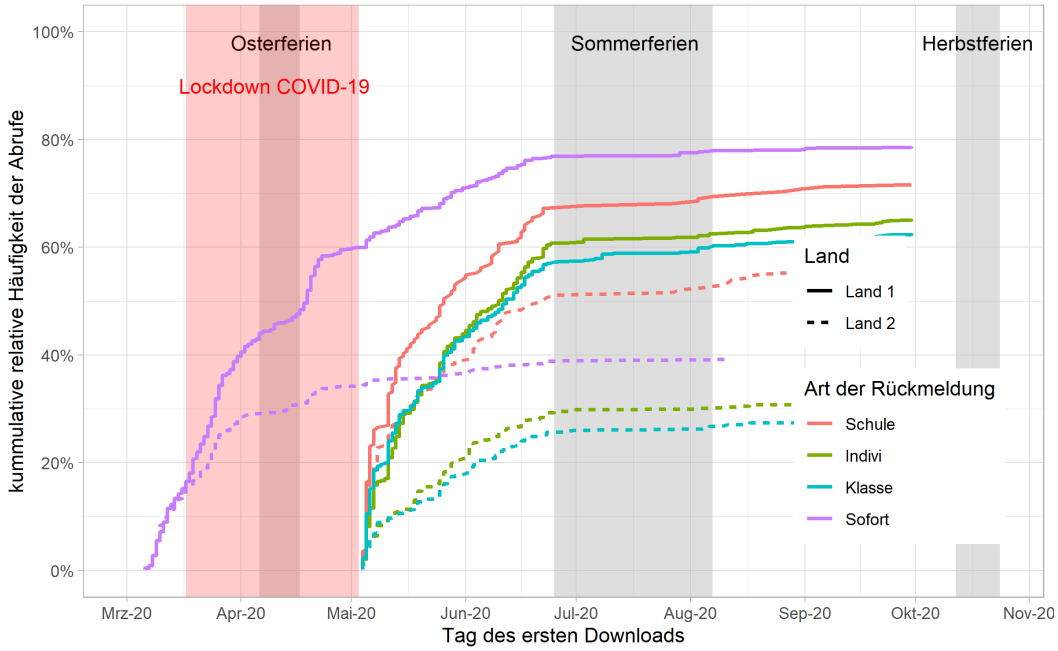


Abbildung A.7.: Downloads von Rückmeldungen für das Fach Deutsch, nach Land

Tabelle A.15.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Deutsch für Berlin und Brandenburg

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-0.4357	0.0699	-6.2323	0.0000	0.6468	0.5636; 0.7414
	Land	1.7429	0.1036	16.8161	0.0000	5.7138	4.6700; 7.0115
	R^2	Cohens d			Deviance	dof	AIC
	0.0395	0.2029			570.54	443	574.54
Indivi	(Intercept)	-0.8055	0.0739	-10.9007	0.0000	0.4469	[0.3861; 0.5159]
	Land	1.4360	0.0989	14.5138	0.0000	4.2037	[3.4664; 5.1091]
	R^2	Cohens d			Deviance	dof	AIC
	0.1129	0.3568			2376.99	1875	2380.99
Klasse	(Intercept)	-0.9691	0.0765	-12.6759	0.0000	0.3794	[0.3260; 0.4400]
	Land	1.4846	0.1002	14.8190	0.0000	4.4134	[3.6310; 5.3780]
	R^2	Cohens d			Deviance	dof	AIC
	0.1185	0.3666			2356.51	1875	2360.51
Schule	(Intercept)	0.2108	0.1359	1.5513	0.1208	1.2347	[0.9468; 1.6142]
	Land	0.7179	0.2007	3.5776	0.0003	2.0501	[1.3867 3.0475]
	R^2	Cohens d			Deviance	dof	AIC
	0.0395	0.2029			570.54	443	574.54

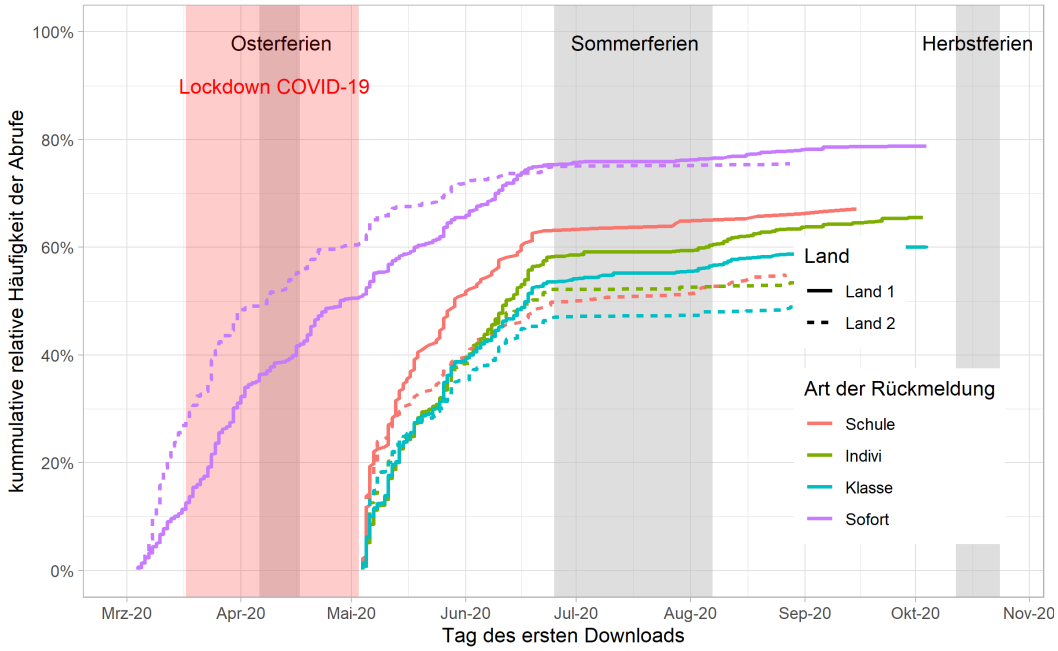


Abbildung A.8.: Downloads von Rückmeldungen für das Fach Englisch, nach Land

Tabelle A.16.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Englisch für Berlin und Brandenburg

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	1.1303	0.0762	14.8381	0.0000	3.0965	[2.6719; 3.6022]
	Land	0.1828	0.1090	1.6779	0.0934	1.2006	[0.9698; 1.4869]
	R^2	Cohens d			Deviance	dof	AIC
	0.0015	0.0383			2056.82	1918	2060.82
Indivi	(Intercept)	0.1502	0.0656	2.2883	0.0221	1.1620	[1.0219; 1.3219]
	Land	0.4962	0.0938	5.2885	0.0000	1.6424	[1.3670; 1.9748]
	R^2	Cohens d			Deviance	dof	AIC
	0.0146	0.1216			2558.59	1918	2562.59
Klasse	(Intercept)	-0.0300	0.0654	-0.4581	0.6469	0.9705	[0.8536; 1.1033]
	Land	0.4456	0.0923	4.8280	0.0000	1.5614	[1.3034; 1.8716]
	R^2	Cohens d			Deviance	dof	AIC
	0.0121	0.1109			2619.80	1918	2623.80
Schule	(Intercept)	0.1942	0.1364	1.4233	0.1546	1.2143	[0.9301; 1.5889]
	Land	0.5193	0.1975	2.6291	0.0086	1.6809	[1.1430; 2.4809]
	R^2	Cohens d			Deviance	dof	AIC
	0.0214	0.1478			580.00	437	584.00

A.7.5. Unterschiede zwischen Ländern und Schulformen

Tabelle A.17.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Mathematik für die zwei Länder sowie verschiedene Schulformen

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-1,4656	0,0980	-14,9497	0,0000	0,2309	0,1898; 0,2788
	Schulform	3,4192	0,1864	18,3409	0,0000	30,5464	21,4408; 44,5889
	Land	2,3795	0,1281	18,5731	0,0000	10,8000	8,4284; 13,9300
	R^2	Cohens d			Deviance	dof	AIC
	0,3852	0,7916			1669,02	1975	1675,02
Indivi	(Intercept)	-1,6254	0,0960	-16,9281	0,0000	0,1968	0,1625; 0,2368
	Schulform	2,2055	0,1252	17,6163	0,0000	9,0745´	7,1249; 11,6420
	Land	1,8060	0,1130	15,9805	0,0000	6,0860	4,8885; 7,6149
	R^2	Cohens d			Deviance	dof	AIC
	0.0146	0.1216			2558.59	1918	2562.59
Klasse	(Intercept)	-1,8174	0,0991	-18,3370	0,0000	0,1624	0,1332 0,1965
	Schulform	2,0254	0,1183	17,1264	0,0000	7,5788	6,0276; 9,5845
	Land	1,7358	0,1120	15,4934	0,0000	5,6735	4,5657; 7,0849
	R^2	Cohens d			Deviance	dof	AIC
	0,2679	0,6050			2110,21	1975	2116,21
Schule	(Intercept)	-0.2419	0.1545	-1.5652	0.1175	0.7851	[0.5787; 1.0617]
	Schulform	0.7553	0.2084	3.6239	0.0003	2.1282	[1.4201; 3.2182]
	Land	0.5973	0.1972	3.0289	0.0025	1.8172	[1.2365; 2.6804]
	R^2	Cohens d			Deviance	dof	AIC
	0.0698	0.2739			582.86	443	588.86

Tabelle A.18.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Deutsch für die zwei Länder sowie verschiedene Schulformen

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	-1.5372	0.1053	-14.5932	0.0000	0.2150	[0.1740; 0.2631]
	Schulform	3.1158	0.1691	18.4240	0.0000	22.5505	[16.3247; 31.6998]
	Land	2.2359	0.1301	17.1875	0.0000	9.3546	[7.2750; 12.1177]
	R^2	Cohens d			Deviance	dof	AIC
	0.3609	0.7515			1675.44	1874	1681.44
Indivi	(Intercept)	-1.6090	0.0993	-16.2107	0.0000	0.2001	[0.1641; 0.2422]
	Schulform	1.9984	0.1206	16.5667	0.0000	7.3775	[5.8415; 9.3754]
	Land	1.6428	0.1131	14.5216	0.0000	5.1697	[4.1510; 6.4692]
	R^2	Cohens d			Deviance	dof	AIC
	0.2542	0.5839			2051.33	1874	2057.33
Klasse	(Intercept)	-1.7521	0.1019	-17.1952	0.0000	0.1734	[0.1415; 0.2110]
	Schulform	1.8815	0.1177	15.9797	0.0000	6.5632	[5.2244; 8.2905]
	Land	1.6767	0.1133	14.7942	0.0000	5.3481	[4.2928; 6.6956]
	R^2	Cohens d			Deviance	dof	AIC
	0.2468	0.5724			2061.30	1874	2067.30
Schule	(Intercept)	-0.1740	0.1564	-1.1128	0.2658	0.8403	[0.6173; 1.1407]
	Schulform	1.1838	0.2295	5.1576	0.0000	3.2666	[2.1017; 5.1783]
	Land	0.7070	0.2073	3.4103	0.0006	2.0280	[1.3540; 3.0548]
	R^2	Cohens d			Deviance	dof	AIC
	0.1239	0.3760			541.37	442	547.37

Tabelle A.19.: Ergebnisse der logistischen Regression zur Untersuchung der Abrufe von Rückmeldungen im Fach Englisch für die zwei Länder sowie verschiedene Schulformen

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Sofort	(Intercept)	0.7878	0.0814	9.6734	0.0000	2.1985	[1.8768; 2.5830]
	Schulform	1.5404	0.1528	10.0815	0.0000	4.6662	[3.4866; 6.3528]
	Land	0.0957	0.1126	0.8500	0.3953	1.1004	[0.8825; 1.3724]
	R^2	Cohens d			Deviance	dof	AIC
	0.0672	0.2685			1926.02	1917	1932.02
Indivi	(Intercept)	-0.1178	0.0730	-1.6137	0.1066	0.8888	[0.7701; 1.0255]
	Schulform	0.9135	0.1056	8.6547	0.0000	2.4932	[2.0303; 3.0715]
	Land	0.4492	0.0958	4.6879	0.0000	1.5670	[1.2990; 1.8914]
	R^2	Cohens d			Deviance	dof	AIC
	0.0544	0.2399			2479.38	1917	2485.38
Klasse	(Intercept)	-0.3323	0.0735	-4.5225	0.0000	0.7173	[0.6208; 0.8280]
	Schulform	1.0045	0.1030	9.7520	0.0000	2.7306	[2.2343; 3.3463]
	Land	0.3935	0.0948	4.1507	0.0000	1.4821	[1.2311; 1.7852]
	R^2	Cohens d			Deviance	dof	AIC
	0.0624	0.2579			2519.58	1917	2525.58
Schule	(Intercept)	-0.1415	0.1554	-0.9100	0.3628	0.8681	[0.6391; 1.1767]
	Schulform	1.0216	0.2196	4.6515	0.0000	2.7776	[1.8179; 4.3066]
	Land	0.4913	0.2028	2.4229	0.0154	1.6344	[1.0998; 2.4370]
	R^2	Cohens d			Deviance	dof	AIC
	0.0897	0.3140			556.92	436	562.92

A.7.6. Unterschiede zwischen Ländern, Schulformen und Leistung

Tabelle A.20.: Ergebnisse der logistischen Regression zur Untersuchung Abhängigkeit der Abrufe von Rückmeldungen im Fach Mathematik vom Land, der Schulform und der Leistung

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Schule	(Intercept)	-0.2114	0.1635	-1.2926	0.1961	0.8095	[0.5863; 1.1142]
	Leistung	-0.1823	0.3217	-0.5668	0.5708	0.8333	[0.4397; 1.5612]
	Schulform	0.9045	0.3363	2.6894	0.0072	2.4708	[1.2844; 4.8259]
	Land	0.5910	0.1975	2.9920	0.0028	1.8058	[1.2279; 2.6652]
	R^2	Cohens d			Deviance	dof	AIC
	0.0707	0.2758			582.54	442	590.54
Indivi	(Intercept)	-1.6145	0.1073	-15.0490	0.0000	0.1990	[0.1606; 0.2446]
	Leistung	-0.0331	0.1470	-0.2251	0.8219	0.9674	[0.7239; 1.2888]
	Schulform	2.2281	0.1608	13.8542	0.0000	9.2825	[6.8018; 12.7810]
	Land	1.8013	0.1149	15.6806	0.0000	6.0573	[4.8482; 7.6075]
	R^2	Cohens d			Deviance	dof	AIC
	0.2873	0.6349			2071.80	1974	2079.80
Klasse	(Intercept)	-1.7806	0.1095	-16.2563	0.0000	0.1685	[0.1354; 0.2081]
	Leistung	-0.1159	0.1503	-0.7710	0.4407	0.8906	[0.6614; 1.1927]
	Schulform	2.1059	0.1585	13.2890	0.0000	8.2145	[6.0487; 11.2623]
	Land	1.7205	0.1136	15.1410	0.0000	5.5871	[4.4823; 6.9991]
	R^2	Cohens d			Deviance	dof	AIC
	0.2682	0.6053			2109.61	1974	2117.61
Sofort	(Intercept)	-1.4340	0.1102	-13.0153	0.0000	0.2384	[0.1912; 0.2946]
	Leistung	-0.0952	0.1540	-0.6180	0.5366	0.9092	[0.6720; 1.2297]
	Schulform	3.4832	0.2137	16.3019	0.0000	32.5651	[21.6378; 50.0564]
	Land	2.3653	0.1299	18.2048	0.0000	10.6476	[8.2810; 13.7842]
	R^2	Cohens d			Deviance	dof	AIC
	0.3853	0.7918			1668.64	1974	1676.64

Tabelle A.21.: Ergebnisse der logistischen Regression zur Untersuchung Abhängigkeit der Abrufe von Rückmeldungen im Fach Deutsch vom Land, der Schulform und der Leistung

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Schule	(Intercept)	-0.2036	0.1629	-1.2501	0.2113	0.8158	[0.5915; 1.1214]
	Leistung	0.2700	0.4120	0.6553	0.5122	1.3100	[0.5876; 2.9974]
	Schulform	0.9415	0.4355	2.1618	0.0306	2.5637	[1.0759; 6.0088]
	Land	0.7147	0.2078	3.4387	0.0006	2.0436	[1.3631; 3.0816]
	R^2	Cohens d			Deviance	dof	AIC
	0.1251	0.3781			540.94	441	548.94
Indivi	(Intercept)	-1.6232	0.1077	-15.0714	0.0000	0.1973	[0.1591; 0.2427]
	Leistung	0.0573	0.1675	0.3423	0.7321	1.0590	[0.7600; 1.4667]
	Schulform	1.9541	0.1767	11.0605	0.0000	7.0577	[5.0158; 10.0314]
	Land	1.6490	0.1146	14.3844	0.0000	5.2017	[4.1648; 6.5291]
	R^2	Cohens d			Deviance	dof	AIC
	0.2543	0.5839			2051.21	1873	2059.21
Klasse	(Intercept)	-1.7921	0.1108	-16.1749	0.0000	0.1666	[0.1335; 0.2062]
	Leistung	0.1601	0.1684	0.9504	0.3419	1.1736	[0.8407; 1.6283]
	Schulform	1.7580	0.1745	10.0741	0.0000	5.8008	[4.1396; 8.2095]
	Land	1.6943	0.1151	14.7239	0.0000	5.4427	[4.3544; 6.8379]
	R^2	Cohens d			Deviance	dof	AIC
	0.2471	0.5729			2060.41	1873	2068.41
Sofort	(Intercept)	-1.5729	0.1155	-13.6154	0.0000	0.2074	[0.1645; 0.2589]
	Leistung	0.1362	0.1754	0.7763	0.4376	1.1459	[0.8119; 1.6160]
	Schulform	3.0136	0.2136	14.1052	0.0000	20.3598	[13.4916; 31.1924]
	Land	2.2537	0.1325	17.0123	0.0000	9.5225	[7.3728; 12.3965]
	R^2	Cohens d			Deviance	dof	AIC
	0.3611	0.7518			1674.84	1873	1682.84

Tabelle A.22.: Ergebnisse der logistischen Regression zur Untersuchung Abhängigkeit der Abrufe von Rückmeldungen im Fach Englisch vom Land, der Schulform und der Leistung

Art	Parameter	est	std.err	z.value	Pr(>|z|)	e^{est}	CI(95%)
Schule	(Intercept)	-0.1656	0.1603	-1.0332	0.3015	0.8474	[0.6178; 1.1594]
	Leistung	0.2534	0.4072	0.6223	0.5337	1.2884	[0.5828; 2.9186]
	Schulform	0.7943	0.4265	1.8621	0.0626	2.2128	[0.9443; 5.0945]
	Land	0.4903	0.2029	2.4169	0.0157	1.6328	[1.0985; 2.4350]
	R^2	Cohens d			Deviance	dof	AIC
	0.0909	0.3161			556.53	435	564.53
Indivi	(Intercept)	-0.1957	0.0815	-2.3999	0.0164	0.8223	[0.7005; 0.9645]
	Leistung	0.2891	0.1337	2.1620	0.0306	1.3353	[1.0281; 1.7371]
	Schulform	0.6950	0.1463	4.7496	0.0000	2.0036	[1.5042; 2.6702]
	Land	0.4716	0.0966	4.8845	0.0000	1.6026	[1.3267; 1.9373]
	R^2	Cohens d			Deviance	dof	AIC
	0.0567	0.2452			2474.68	1916	2482.68
Klasse	(Intercept)	-0.3656	0.0818	-4.4689	0.0000	0.6937	[0.5906; 0.8140]
	Leistung	0.1238	0.1327	0.9336	0.3505	1.1318	[0.8723; 1.4678]
	Schulform	0.9110	0.1436	6.3438	0.0000	2.4868	[1.8783; 3.2989]
	Land	0.4027	0.0954	4.2230	0.0000	1.4959	[1.2411; 1.8038]
	R^2	Cohens d			Deviance	dof	AIC
	0.0628	0.2589			2518.71	1916	2526.71
Sofort	(Intercept)	0.7621	0.0902	8.4458	0.0000	2.1427	[1.7980; 2.5615]
	Leistung	0.0954	0.1461	0.6527	0.5140	1.1001	[0.8283; 1.4694]
	Schulform	1.4679	0.1891	7.7609	0.0000	4.3399	[3.0058; 6.3161]
	Land	0.1035	0.1132	0.9139	0.3607	1.1090	[0.8883; 1.3848]
	R^2	Cohens d			Deviance	dof	AIC
	0.0674	0.2689			1925.59	1916	1933.59

Selbstständigkeitserklärung

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten Arbeit um eine von mir selbstständig und ohne fremde Hilfe verfasste Arbeit handelt.

Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit verwendeten fremden Quellen, auch aus dem Internet (einschließlich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken o. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken o.Ä. anderer Autorinnen und Autoren die Quelle angegeben habe.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend geahndet werden.

17. November 2021 Peter Harych