Institut zur Qualitätsentwicklung im Bildungswesen

Schlusssachbericht des Projekts „Technologiebasiertes Assessment (TBA) – Machbarkeitsstudie“

Das Projekt „Technologiebasiertes Assessment (TBA) – Machbarkeitsstudie“ wurde von 2018 bis 2020 am IQB durchgeführt. Der Schlusssachbericht wird in zwei Teilen vorgelegt: Teil A beschreibt die Ergebnisse des Zeitraums 2018-2019 und Teil B die Ergebnisse des Jahres 2020.

Teil A: Ergebnisse der TBA-Machbarkeitsstudie 2018-2019

Die „Vereinbarung zur Weiterentwicklung der Vergleichsarbeiten (VERA)“ (Beschluss der Kultusministerkonferenz vom 08.03.2012 i. d. F. vom 15.03.2018) sieht neben einer Modularisierung und Flexibilisierung auch die Bereitstellung der VERA-Tests als Onlineinstrument vor. Demnach soll „die Durchführung von VERA-8 nach fachlichen und administrativen Möglichkeiten mittelfristig“ sowie „perspektivisch auch für VERA-3“ (S. 6) auf technologiebasiertes Testen (TBA) umgestellt werden.

Das IQB führte seit 01.01.2018 eine TBA-Machbarkeitsstudie (Laufzeitende 31.12.2019) durch, in der untersucht wurde,

(1) wie groß der Aufwand sein wird, der mit einer Umstellung auf TBA verbunden ist (u. a. Funktionsumfang des Testsystems, technische Mindestvoraussetzungen, Kosten);

(2) inwieweit Moduseffekte auftreten bzw. die Ergebnisse einer computerbasierten Testung mit denen einer papierbasierten Testung vergleichbar sind und weiterhin auf der vorliegenden Bildungsstandardmetrik abgebildet werden können sowie

(3) welche Konsequenzen mit der Entwicklung und dem Einsatz innovativer Aufgabenformaten verbunden sind (u. a. Erfahrungen aus dem Aufgabenentwicklungsprozess, Konstruktäquivalenz innovativer und regulärer Aufgabenformate).

Um die genannten Fragestellungen zu untersuchen, hat das IQB seit Projektbeginn mehrere interne Erprobungen am IQB, zwei technische Erprobungsstudien mit regulären und innovativen Aufgabenformaten an Schulen in Berlin, Brandenburg und Nordrhein-Westfalen sowie eine an die Pilotierung zum IQB-Bildungstrend 2021 angekoppelte umfangreiche Machbarkeitsstudie mit 2676 Schülerinnen und Schülern der 9. Klassenstufe im Fach Englisch für die Kompetenzbereiche Lese- und Hörverstehen durchgeführt.

Nachfolgend werden insbesondere die Ergebnisse zu den Fragestellungen (1) und (2) dargestellt und Empfehlungen für den Umgang mit Moduseffekten im Rahmen von VERA sowie Konsequenzen abgeleitet, die bei einer Umstellung auf TBA zu berücksichtigen sind. Entsprechend des Vorschlags zur Sicherstellung der Anschlussfähigkeit des TBA-Projekts (Stand: 14.03.2019) werden Ergebnisse zu Fragestellung (3) primär in Teil B berichtet.

1.1 Aufwand bei der Umstellung auf TBA

1.1.1 Welches Testsystem eignet sich für eine onlinebasierte Testdurchführung mit Schulhardware?

Damit das technische System zukünftig vom IQB für die Aufgabenentwicklung, Testdurchführung und Datenaufbereitung verwendet und von den Ländern sowohl komplett als eigenständige Anwendung eingesetzt als auch bausteinartig in vorhandene Softwarelösungen integriert werden kann, sollte es als Open-Source-Version verfügbar und kostenfrei nutzbar sein. In der Machbarkeitsstudie wurden deshalb zunächst die infrage kommenden frei verfügbaren Open-Source-Systeme Cap3, TAO (Basisversion) und ItemBuilder (DIPF) auf ihre Eignung geprüft und erprobt. Dabei zeigte sich, dass diese Testsysteme die genannten Anforderungen in der vorliegenden Version unter anderem aus den folgenden zentralen Gründen nicht erfüllen:

- **Cap3:** Es werden nur Standarditemformate unterstützt, die nicht erweiterbar sind, was perspektivisch zu Einschränkungen in der Testentwicklung führen oder den Einkauf kostenintensiver Zusatzprogrammierungen erforderlich machen würde. Zudem können Tests nur auf dem Server des Entwicklers bereitgestellt werden. Für die Durchführung von VERA stellt dies ein Ausschlusskriterium dar, da die Tests in vielen Ländern (u. a. aus datenschutzrechtlichen Gründen) auf landeseigenen Servern bereitgestellt werden müssen.
- **TAO (Basisversion):** Die freie TAO-Basisversion verfügt über eine eingeschränkte Funktionalität (nur Standarditemformate) und Leistungsfähigkeit. Es wären fortlaufend sehr aufwendige Ergänzungsprogrammierungen erforderlich um die Nutzbarkeit aufrecht zu erhalten, da regelmäßig neue Programmversionen veröffentlicht werden, an die die verwendete Systemversion sowie ggf. eigene Ergänzungsprogrammierungen in den Ländern angepasst werden müssten (z. B. werden in Italien und Frankreich für die Anwendungen des TAO-Systems dauerhaft umfangreiche Dienstleistungen der Firma OAT eingekauft).
- **ItemBuilder (DIPF):** Das System wird als Offlinevariante u. a. vom NEPS verwendet, es eignet sich derzeit aber noch nicht für Onlinetestungen mit hohen Schülerzugriffszahlen wie in VERA (vgl. u. a. Ergebnisse der technischen Erprobung innovativer Aufgabenformate, Abschnitt 1.3). Zudem ist die Handhabung des Testsystems bei der Erstellung und Bearbeitung von Aufgaben derzeit noch zu aufwendig, da umfassende Programmierkenntnisse benötigt werden.

Als Konsequenz aus der Erprobung bereits vorhandener Testsysteme, mit denen sowohl die Durchführung der Machbarkeitsstudie als auch eine dauerhafte Nutzung für Testungen des IQB nur mit Einschränkungen möglich wären, wurde ein eigenes onlinebasiertes Open-Source-Testsystem ent-

wickelt. Bei allen Entwicklungsschritten waren und sind zwei Aspekte maßgeblich: Zum einen muss das Testsystem sowohl vom IQB als auch von Dritten (insbesondere den Ländern bzw. auswertenden Einrichtungen) dauerhaft genutzt und erweitert bzw. an länderspezifische Besonderheiten angepasst werden können. Dazu müssen der Programmcode frei verfügbar und die Systemarchitektur im Sinne eines Baukastenprinzips konzipiert sein, um in bereits bestehende technische Systeme integriert werden zu können. Um dies zu gewährleisten, wurde mit Projektbeginn die Fachgruppe „VERA Online“ eingerichtet, in der das IQB gemeinsam mit Vertreterinnen und Vertretern der Länder bzw. auswertenden Einrichtungen u. a. Schnittstellen definiert. Zum anderen zielen alle Entwicklungsschritte darauf ab, die technischen Mindestanforderungen an das Testsystem und die Nutzerinnen und Nutzer möglichst niedrig zu halten, damit es mit der verfügbaren Schultechnik eingesetzt und sowohl von Lehrkräften als auch von Schülerinnen und Schülern ohne besondere Computerkenntnisse intuitiv bedient werden kann.

Um mit dem IQB-Testsystem Kompetenztests online durchzuführen, wurde in der Projektlaufzeit begonnen, folgende Funktionsbereiche zu implementieren:

- *Online-Aufgabenentwicklung und -erstellung*: Für das Anlegen von Testaufgaben wurde ein Autorenmodul mit grundlegenden Funktionen entwickelt, mit denen Aufgabenentwicklerinnen und -entwickler Testaufgaben über einen Editor gestalten und Auswertungsvorschriften anlegen können. Derzeit verfügt das Autorenmodul über die Funktionen und Aufgabenformate, die in der Machbarkeitsstudie benötigt wurden. Bis zum Ende der aktuellen Projektlaufzeit (31.12.2020) sollen der Funktionsumfang, die Bedienbarkeit und die Bandbreite unterstützter Aufgabenformate erweitert werden, indem beispielsweise auch die Verwendung von Hilfsmitteln (zunächst standardisierter Taschenrechner) bei der Aufgabenbearbeitung ermöglicht wird. Ferner ist es für das IQB wesentlich, eine große Anzahl von Aufgaben effizient, zuverlässig und nachhaltig erstellen, pflegen und dokumentieren zu können. Die dafür erforderlichen Funktionsbereiche müssen weiter ausgebaut werden (z. B. durch die Anbindung eines Metadatenkatalogs, und der Itemdatenbank).
- *Online-Testdurchführung:* Für die Durchführung von Kompetenztests in Schulen wurde in der Projektlaufzeit begonnen, eine Testleiterkonsole zu entwickeln, die einen grundlegenden Systemcheck vor dem Test, die Verwaltung der Testaufgaben und Testpersonen, das Beobachten der Testdurchführung sowie das Herunterladen der Testantworten im Rohformat ermöglicht. Diese Funktionen erfordern keine besonderen Vorkenntnisse durch die Testleitung (z. B. Lehrkräfte).
- *Online-Datenverarbeitung und Datenaufbereitung:* Da die Schülerantworten zunächst im Rohformat vorliegen und für die Analyse aufbereitet werden müssen, wurde für die Machbarkeitsstudie eine Offline-Anwendung programmiert, mit der Excel-Dateien erzeugt werden können. Der Datenaufbereitungsprozess ist momentan noch aufwendig. Perspektivisch soll mit Hilfe eines Antwortkonverters ein Primärdatensatz erzeugt werden, der von den auswertenden Einrichtungen in den Ländern direkt ohne Zwischenschritte weiterverarbeitet werden kann.

- *Technologische Infrastruktur:* Um den Austausch von Systemkomponenten zwischen dem IQB und den Ländern zuverlässig und nachhaltig zu gestalten, wurde in der Projektlaufzeit begonnen, dafür erforderliche Dienste und Dokumentationen zu entwickeln (u. a. Metadatenkatalog, Code-Verwaltung, Plattform für Erfahrungsaustausch).

Die Qualitätssicherung der beschriebenen Entwicklungsarbeiten erfolgt im laufenden Projektzeitraum zum einen durch externe Beratungen, die zur Optimierung des Programmcodes und zur Sicherstellung einer hohen Leistungsfähigkeit sowie Anpassbarkeit an sich verändernde technische und infrastrukturelle Bedingungen erforderlich sind. Zum anderen wird durch den regelmäßigen Austausch in der länderübergreifenden Fachgruppe „VERA Online" sichergestellt, dass die Anforderungen der Länder in den Entwicklungsprozess einfließen und die Anschlussfähigkeit an die Arbeitsstrukturen und an die technischen Systeme in den Ländern und auswertenden Einrichtungen gegeben ist.

In der Machbarkeitsstudie wurden durch das IQB weitreichende Vorarbeiten für die Entwicklung der genannten Online-Anwendungen geleistet. In der aktuellen Version hat sich das Testsystem in verschiedenen Erprobungen am IQB, einer technischen Erprobung (Oktober 2018; 12 Schulen, $N = 251$) und in der Machbarkeitsstudie (Frühjahr 2019; 123 Klassen, $N = 2676$) bewährt. Bis zum Abschluss des laufenden Projekts (31.12.2020) werden grundlegende Basisfunktionen für die Aufgabenerstellung, Testdurchführung und Datenaufbereitung zur Verfügung stehen, sodass Testdurchführungen mit dem IQB-Testsystem ab dem Jahr 2020 möglich sein werden.

Das IQB empfiehlt, die derzeit noch weitgehend rudimentäre Basisversion des IQB-Testsystems in den kommenden Jahren weiter zu optimieren und um wichtige Funktionsbereiche zu erweitern, um eine Standardversion des Testsystems bereitstellen zu können. Die Weiterentwicklung der Qualität und Leistungsfähigkeit der oben genannten Funktionsbereiche ist erforderlich, damit das System sowohl von auswertenden Einrichtungen als auch von Lehrkräften und Schülerinnen und Schülern intuitiv bedient werden kann. Insbesondere bei Lehrkräften dürfte eine niedrigschwellige und komfortable computerbasierte Testdurchführung ohne besondere Vorkenntnisse die Akzeptanz von VERA steigern.

1.1.2 Welche technischen Mindestvoraussetzungen werden in den teilnehmenden Schulen benötigt, damit das IQB-Testsystem für eine Onlinetestung genutzt werden kann?

Zur Durchführung der Machbarkeitsstudie wurde für die Schülerinnen und Schüler je ein Computer bzw. Laptop benötigt, der mit Bildschirm, Tastatur, Maus, Audioausgang (einschließlich Kopfhörer) sowie Internetverbindung und einer aktuellen Browserversion von Firefox oder Chrome ausgestattet war. Alternativ konnte eine aktuelle, portable Browserversion genutzt werden, die auf einem USB-Stick bereitgestellt wurde. Auf Basis eines Systemchecks im Vorfeld der Machbarkeitsstudie, an dem nur Schulen teilnahmen, die angaben, über Computer an ihrer Schule zu verfügen, erfüllten 160 von 169 der allgemeinbildenden Schulen (95 %) und 51 von 57 Förderschulen (89 %) diese Voraussetzungen[1]. Während der Testbearbeitung traten nur in Einzelfällen technische Schwierigkeiten auf

[1] Eine genaue Beschreibung der Stichprobenziehung in der Machbarkeitsstudie kann Abschnitt 1.2.2 entnommen werden.

(Probleme beim Abspielen von Audiodateien: 4 %, defekte Kopfhörer: 1 %, Navigationsschwierigkeiten beim Weiterblättern: 1 %, PC-Absturz: 1 %).

Insgesamt weisen die Ergebnisse darauf hin, dass die technischen Mindestvoraussetzungen in der Mehrheit der Schulen, die über eine Computerausstattung verfügen, erfüllt werden. Zudem traten nur sehr vereinzelt technische Probleme während der Testbearbeitung auf. Die Basisversion des IQB-Testsystems hat sich damit insgesamt als zuverlässig erwiesen.

1.1.3 Wie sollten Aufgaben im TBA-Format auf dem Bildschirm dargestellt werden?

Empirische Studien zu Moduseffekten weisen darauf hin, dass die Darstellung von Aufgaben im TBA-Format (z. B. einseitige Darstellung vs. mehrseitige Darstellung der Items zu einem Stimulus) die Lösungswahrscheinlichkeit und somit die Schwierigkeit der Items beeinflussen kann (Kröhne & Martens, 2011).

In der Machbarkeitsstudie des IQB wurde bei der Übertragung der papierbasierten Aufgaben in das computerbasierte Format darauf geachtet, überwiegend Darstellungsvarianten umzusetzen, die möglichst vergleichbare Durchführungsbedingungen gewährleisten wie bei einer Testbearbeitung im Papierformat. So ist es beispielsweise wichtig, dass ein Stimulustext wie in einem papierbasierten Testheft auch bei der Bearbeitung am Computer permanent sichtbar ist, da es sonst zu Einschränkungen in der Konstruktvalidität kommen kann, weil beispielsweise eine stärker gedächtnisbasierte Leseleistung erfasst wird. Gleichzeitig wurde an einigen Stellen bewusst vom papierbasierten Testformat abgewichen, um die Potenziale von TBA auszuschöpfen und die Durchführungsvalidität im Vergleich zum Papierformat zu erhöhen. Dies gilt beispielsweise für das Abspielen von Zuhörtexten, die in einer papierbasierten Testung bisher von der Lehrkraft über einen CD-Player für alle Schülerinnen und Schüler in der Klasse gleichzeitig abgespielt werden. Störende Nebengeräusche lassen sich dabei nicht vermeiden und ein individuelles Abspielen ist ebenfalls nicht möglich. Im IQB-Testsystem können abspielbare Mediendateien (z. B. Zuhörtexte) von den Schülerinnen und Schülern hingegen individuell über Kopfhörer in angepasster Lautstärke abgespielt werden.

Das IQB strebt an, den Ländern die TBA-Aufgaben zukünftig ebenso wie die papierbasierten Aufgaben in einem standardisierten Format zu übergeben. Unter Berücksichtigung von Erkenntnissen aus vorliegenden Moduseffektstudien (zusammenfassend vgl. Kröhne & Martens, 2011) wurden für die Machbarkeitsstudie folgende Darstellungsstandards bei der Itemdarstellung implementiert:

- Eine Testaufgabe, die aus einem Stimulus (z. B. Lesetext, Hörtext) und dazugehörigen Items besteht, wird in der Regel in einem vertikal geteilten Bildschirm dargestellt (Splitscreen)[2]. Im linken Bildschirmfenster befindet sich der Stimulus, im rechten Bildschirmfenster sind die Items dargestellt. Beide Bildschirmfenster lassen sich unabhängig voneinander scrollen.

[2] Ausnahmen sind sehr kurze Aufgaben und innovative Aufgabenformate, die teilweise eine andere Bildschirmaufteilung erfordern, weil zusätzliche Anwendungen (z. B. dynamische Geometriesoftware, Tabellenkalkulationsprogramm) genutzt werden.

- Alle Items eines Stimulus werden untereinander dargestellt, d. h. die Schülerinnen und Schüler müssen bei vielen Items vertikal scrollen. Dies hat sich im Vergleich zur Darstellung einzelner Items auf mehreren Bildschirmseiten, zwischen denen geblättert werden muss, als intuitiver bedienbar erwiesen (siehe Abschnitt 1.2.3, Tabelle 1.1).
- Antwortalternativen von Multiple-Choice- oder Forced-Choice-Items werden untereinander dargestellt.
- Die Aufgaben werden für eine optimale Darstellung mit einer Bildschirmauflösung von mindestens 1280 x 1024 Pixel angelegt.
- Der Aufgabentitel wird in der Kopfzeile angezeigt. Für Aufgabeninstruktionen und Items wird die Schriftart Arial (Instruktionstext: 24pt, Itemtext: 20pt) verwendet, für den Stimulus die Schriftart Times New Roman (24pt).
- Mediendateien (z. B. Audio, Video) können individuell gestartet werden. Um eine standardisierte Durchführung sicherzustellen, kann die weitere Steuerung von Mediendateien (Pausieren, Wiederholen) reguliert werden. In der Machbarkeitsstudie wurde die Zuhör-Instruktion automatisch beim ersten Aufrufen einer Hörverständnisaufgabe gestartet. Anschließend konnten die Schülerinnen und Schüler den Zuhörtext individuell starten, jedoch während des Zuhörens nicht pausieren oder beliebig häufig wiederholen.
- Auf jeder Bildschirmseite sind unsichtbare „Viewpoints“ platziert, die von den Schülerinnen und Schülern angesehen werden müssen und teilweise vertikales Scrollen erforderlich machen, bevor die Bearbeitung der nächsten Aufgabe möglich ist.
- In der Kopfzeile des Testfensters ist die Anzahl insgesamt zu bearbeitender Testaufgaben zu erkennen und markiert, wie viele und welche Aufgaben bereits bearbeitet wurden. Fünf Minuten vor dem Bearbeitungsende wird ein Popupfenster eingeblendet, das auf die verbleibende Testzeit hinweist. Zur Sicherstellung der Durchführungsvalidität wurde die Bearbeitung des Lese- bzw. Hörverstehenstestteils durch ein kurzes, einsilbiges Freigabewort synchronisiert, das in ein Eingabefeld eintragen werden musste, um mit der Testbearbeitung fortzufahren.

  Diese Darstellungsstandards sollten im Austausch mit den Ländern bzw. auswertenden Einrichtungen sowie nationalen und internationalen Kooperationspartnern (z. B. DIPF, LUCET) zukünftig weiter erprobt und erweitert werden.

1.1.4 Wie hoch sind die mittelfristigen am IQB anfallenden Kosten bei einer Umstellung auf TBA?

Bei einer Umstellung der Testinstrumente des IQB auf TBA muss das IQB-Testsystem in die vorhandenen Arbeitsstrukturen des IQB integriert werden. Das Testsystem würde in diesem Fall zukünftig in allen Prozesse der Aufgabenentwicklung (z. B. Erstellen und Überarbeiten von Aufgabenentwürfen durch interne Mitarbeitende und externe Kooperationspartner, Länderreview) sowie der Vorbereitung, Durchführung und Auswertung von Tests eingesetzt werden und muss dafür um Funktionen erweitert werden, die eine reibungslose Integration in vorhandene Systemlösungen (z. B. IQB-Itemdatenbank, Austauschplattformen, Datenaufbereitungs- und -auswertungsroutinen) gewährleis-

ten. Dafür werden in den kommenden Jahren bis 2023 zunächst zusätzliche Mittel benötigt, um die Grundlagen für eine effiziente und nachhaltige Nutzung des Testsystems zu schaffen und alle Arbeitsprozesse auf TBA umzustellen. Auf diese Weise wird gewährleistet, dass auch langfristig punktuelle Erweiterungsprogrammierungen (z. B. aufgrund eines veränderten Nutzungsverhaltens) mit vertretbarem Aufwand geleistet werden können.

1.2 Moduseffekte bei der Bearbeitung von papierbasierten und computerbasierten Testaufgaben

Die Umstellung papierbasierter (PPA) auf technologiebasierte Assessments (TBA) geht mit zahlreichen Vorteilen einher. Zu diesen Vorteilen zählen neben der Anschlussfähigkeit an internationale und nationale Studien[3] unter anderem eine Erhöhung der Messgenauigkeit, Entlastung und Akzeptanzsteigerung bei Lehrkräften, Ressourcenentlastung durch den Wegfall von Druck und Versand der Testmaterialien, verbesserte Administrationsbedingungen im Rahmen der VERA-Modularisierung (z. B. perspektivisch durch adaptives bzw. Multistage-Testen auf der Ebene der Ergänzungsmodule[4]), Erhöhung der Konstruktvalidität (z. B. aufgrund der breiteren inhaltlichen Abdeckung der Bildungsstandards durch die Nutzung innovativer Aufgabenformate, siehe Abschnitt 1.3) oder Sofortrückmeldungen auf Basis automatisiert auswertbarer geschlossener und halboffener Antworten.

Gleichzeitig sind mit einer Umstellung der Administrationsbedingungen Herausforderungen für die Datenauswertung und Ergebnisrückmeldung verbunden. Insbesondere stellen sich Fragen der Konstruktvalidität, also ob die Konstrukte unabhängig vom Modus (PPA oder TBA) in gleicher Weise gemessen werden können. Sogenannte Moduseffekte geben an, inwieweit Personen die Bearbeitung von Testaufgaben am Computer im Vergleich zu einer papierbasierten Bearbeitung leichter oder schwerer fällt. Die bisherige Forschungslage zur Untersuchung von Moduseffekten ist insgesamt uneinheitlich. Während einige Studien keine bedeutsamen Unterschiede zwischen papier- und computerbasierten Testergebnissen feststellen (Margolin, Driscoll, Toland & Kegler, 2013; Singer & Alexander, 2017; Wang, Jiao, Young, Brooks & Olson, 2008), weisen andere Arbeiten sowohl für Schülerinnen und Schüler in der Primarstufe als auch in der Sekundarstufe I auf Moduseffekte hin, die teilweise mit weiteren Hintergrundmerkmalen der Schülerinnen und Schüler (z. B. Geschlecht, Schulart, soziökonomischer Hintergrund, computerbezogene Kompetenzen) oder der Testgestaltung (z. B. Testdesign, Aufgabenanordnung, Eingabemodalität, Endgerät) interagieren (Chen, Cheng, Chang, Zheng & Huang, 2014; Jerrim, Micklewright, Heine, Sälzer & McKeown, 2018; Kröhne & Martens, 2011).

Die Moduseffektstudie im Kontext der Umstellung von TIMSS auf eTIMSS weist für Schülerinnen und Schüler der 4. Jahrgansstufe sowohl für das Fach Mathematik als auch für die Naturwissenschaften darauf hin, dass den Kindern die Testbearbeitung auf dem Tablet schwerer fiel als die Bearbeitung

[3] So werden etwa PISA, NEPS sowie nationale Assessments in der Schweiz oder in Luxemburg für Schülerinnen und Schüler der Sekundarstufe I bereits seit mehreren Jahren computerbasiert durchgeführt und auch in der Primarstufe kommen in eTIMSS und ePIRLS ab 2019 bzw. 2021 computerbasierte Tests zum Einsatz.

[4] Durch eine restringierte Testheftzusammenstellung (z. B. Content Balancing für alle Leitideen in Mathematik oder Exposure Control, um eine maximal hohe Aufgabennutzung innerhalb einer Klasse zu ermöglichen) kann auch bei einer adaptiven Testgestaltung weiterhin gewährleistet werden, dass die Ergebnisse sinnvoll auf Klassenebene ausgewertet und für die Unterrichtsentwicklung genutzt werden können.

des Tests in Papierform (Fishbein, Martin, Mullis & Foy, 2018). Auch die PISA-Feldtestdaten aus dem Jahr 2014 liefern Hinweise darauf, dass die computerbasierte Testbearbeitung in den Naturwissenschaften, in Mathematik und im Lesen für 15-jährige Jugendliche schwieriger war als die papierbasierte Testbearbeitung (Jerrim et al., 2018; Robitzsch et al., 2016). Die Ergebnisse von PISA 2015 zeigen zudem, dass es insbesondere im Bereich Lesen zusätzliche Interaktionseffekte mit der Geschlechtszugehörigkeit geben könnte, da sich der Leistungsvorsprung der Mädchen im Vergleich zu PISA 2012 gegenüber den Jungen mehr als halbierte (Weis, Zehner, Strohmaier, Artelt & Pfost, 2016). Die Ergebnisse von PISA 2018 weisen erneut auf Moduseffekte in allen drei Kompetenzbereichen hin, wobei diese in den Naturwissenschaften auch für Items beobachtet wurden, die als modusinvariant klassifiziert und für die Fortführung des Trends verwendet werden (Goldhammer et al., 2019). Die Autoren weisen deshalb darauf hin, dass Veränderungen im Trend nur mit Vorsicht interpretiert werden sollten.

Für die vom Projekt „kompetenztest.de“ an der Uni Jena durchgeführten computerbasierten VERA-Erhebungen deuten die Ergebnisse im Schuljahr 2017/18 im Fach Englisch für die Kompetenzbereiche Lese- und Hörverstehen ebenfalls darauf hin, dass den Schülerinnen und Schülern die Testbearbeitung am Computer insgesamt schwerer fällt als die papierbasierte Testbearbeitung und diese Moduseffekte bei Jugendlichen am Gymnasium stärker ausfallen als an nichtgymnasialen Schulen. Jugendliche nichtgymnasialer Schularten erreichten im Hörverstehen im Fach Englisch am Computer tendenziell sogar etwas bessere Kompetenzwerte als bei der Testbearbeitung im Papierformat (Nachtigall, 2018). Fallen Moduseffekte wie in diesem Fall nicht konsistent über Subgruppen der Gesamtpopulation aus, ist eine Ergebnisrückmeldung auf der vorliegenden Bildungsstandardmetrik und Kompetenzstufenmodellen nicht ohne weiteres möglich.

Zusammenfassend lässt sich festhalten, dass die Richtung und Stärke von Moduseffekten von verschiedenen Faktoren abzuhängen scheint, wie etwa vom Kompetenzbereich (Kingston, 2008), von Merkmalen der Testadministration (z. B. Darstellungsvarianten; Bennett et al., 2008; Wang et al., 2008) oder von Hintergrundmerkmalen der Schülerschaft (z. B. Geschlecht; Jerrim et al., 2018). Kröhne und Martens (2011) empfehlen daher, in jeder Studie eine eigene empirische Überprüfung von Moduseffekten vorzunehmen, insbesondere, wenn Unklarheiten über das Zusammenwirken der genannten Faktoren bestehen.

1.2.1 Fragestellungen zur Prüfung von Moduseffekten in der TBA-Machbarkeitsstudie

In der Machbarkeitsstudie des IQB wurde der Frage nachgegangen, inwieweit Moduseffekte im Fach Englisch in den Kompetenzbereichen Lese- und Hörverstehen auftreten. Aus den Befunden ergeben sich Konsequenzen für die Abbildung der Ergebnisse aus bildungsstandardbasierten Tests auf der Metrik der vorliegenden Bildungsstandards und Kompetenzstufenmodelle. Folgende Szenarien sind dabei möglich:

- **Keine Moduseffekte**: Die Ergebnisse aus TBA-basierten Tests können auf der bereits etablierten Bildungsstandardmetrik abgebildet und in Bezug zu den bereits vorliegenden Kompetenzstufenmodellen gesetzt werden.

- **Konsistente Moduseffekte:** Es treten Moduseffekte auf, die für alle Aufgabenformate und in allen Teilpopulationen *vergleichbar* ausfallen. Die Ergebnisse von TBA-basierten Tests können auf der bereits etablierten Bildungsstandardmetrik abgebildet und in Bezug zu den bereits vorliegenden Kompetenzstufenmodellen gesetzt werden, wenn die Item- und Personenparameter mit einer einheitlichen Verrechnungsvorschrift um den Moduseffekt adjustiert werden.
- **Differenzielle Moduseffekte:** Es treten Moduseffekte auf, die sich zwischen Aufgabenformaten und/oder Teilpopulationen *unterscheiden* und somit nicht generalisiert werden können. Die Ergebnisse von TBA-basierten Tests können nicht ohne weiteres auf der bereits etablierten Bildungsstandardmetrik abgebildet und in Bezug zu den bereits vorliegenden Kompetenzstufenmodellen gesetzt werden. Eine Umstellung auf TBA erfordert eine Neunormierung der Bildungsstandards und eine Überarbeitung der Kompetenzstufenmodelle.

1.2.2 Methode

Stichprobe

Die Analysen zur Prüfung von Moduseffekten basieren auf Daten von insgesamt 2676 Neuntklässlerinnen und Neuntklässlern an allgemeinbildenden Schulen[5] (123 Schulen, davon 34.9 % Gymnasien; mittleres Alter der Schülerinnen und Schüler: 15.59 Jahre [*SD* = 0.69]; 48.23 % weiblich; 16.33 % nicht-deutscher Sprachhintergrund[6], 1.87 % mit sonderpädagogischem Förderbedarf), die im Jahr Frühjahr 2019 an der Pilotierung zum Bildungstrend 2021 in der Sekundarstufe I teilnahmen.

Die Stichprobenziehung richtete sich nach dem Königsteiner Schlüssel, wobei neun von 16 Ländern an der Studie teilnahmen, die überwiegend nicht an der Pilotierung zum IQB-Bildungstrend 2018 im Jahr 2017 beteiligt waren. Die Länder übermittelten dem IQB Schullisten, woraufhin Schulen innerhalb der Strata „Gymnasium“, „nichtgymnasiale Schule“ und „Förderschule“ zufällig gezogen wurden. In den Gymnasien und in den nichtgymnasialen Schulen wurde je eine Klasse zufällig ausgewählt; in Förderschulen wurde der gesamte Jahrgang getestet. Die Zuordnung der gezogenen Klassen zur Teilstichprobe, die den Englischtest im papier- und computerbasierten Format bearbeitete, fand zunächst unter Berücksichtigung der Schularten und Teilstichprobengrößen zufällig statt.[7] Auf Basis der Ergebnisse aus dem Systemcheck wurde diese Zuordnung anschließend noch einmal modifiziert, um sicherzustellen, dass in den teilnehmenden Klassen ausreichend geeignete Computer für die TBA-Testung zur Verfügung standen (vgl. Abschnitt 1.1.2). Da nur neun Länder auf Basis des Königsteiner Schlüssels an die Studie einbezogen wurden und eine Anpassung der Teilnahme der

[5] Es nahmen auch 10 Förderschulen an der Machbarkeitsstudie teil. Die Ergebnisse zeigen, dass computerbasierte Tests auch an diesen Schulen zuverlässig durchführbar sind. Da Schülerinnen und Schüler an Förderschulen in der Machbarkeitsstudie nur den TBA-Testteil und nicht den papierbasierten Testteil bearbeiteten, werden sie in die Analysen zur Prüfung von Moduseffekten nicht einbezogen.

[6] Der Sprachhintergrund wurde über die zu Hause gesprochene Umgangssprache erfasst. Jugendliche, die angaben, zu Hause immer oder meistens eine andere Sprache zu sprechen, werden der Gruppe mit nicht-deutscher Herkunftssprache zugeordnet. Jugendliche, die angaben, zu Hause immer oder meistens Deutsch zu sprechen, werden der Gruppe mit deutscher Herkunftssprache zugeordnet.

[77] Teildesign 1: Deutsch, PPA, allgemeinbildende Schulen; Teildesign 2: Englisch, TBA und PPA, allgemeinbildende Schulen; Teildesign 3: Deutsch, PPA und Englisch, TBA, Förderschulen. Die vorliegenden Ergebnisse beruhen auf Teildesign 2.

Klassen an der papier- und computerbasierten Testung nach dem Systemcheck erforderlich war, erfolgte keine Bestimmung von Klassengewichten und Auswertungen auf Populationsebene. Die untersuchte Stichprobe ist somit nur eingeschränkt repräsentativ.

Durchführung

Die Schülerinnen und Schüler bearbeiteten insgesamt sechs Aufgabenblöcke mit einer Bearbeitungszeit von jeweils 20 Minuten (insgesamt jeweils 60 Minuten PPA bzw. TBA). Etwa die Hälfte der teilnehmenden Klassen bearbeitete zuerst den PPA-Testteil und anschließend den TBA-Testteil oder umgekehrt. Zwischen beiden Testteilen lag eine zwanzigminütige Pause. In 62 % der Klassen war die Anzahl zu testender Schülerinnen und Schüler größer als die Anzahl vorhandener Computerarbeitsplätze in einem Raum, sodass die Testgruppe geteilt werden musste. Bei geteilten Testgruppen bearbeitete die erste Gruppe die erste Testhälfte (drei Aufgabenblöcke) im PPA-Modus, während die zweite Testgruppe parallel dazu den ersten Testteil im TBA-Modus in einem anderen Raum absolvierte. In der zwanzigminütigen Pause tauschten beide Testgruppen die Räume und bearbeiteten die zweite Testhälfte im jeweils anderen Modus.

Um die Konfundierung der Ergebnisse mit Positionseffekten zu minimieren, wurde ein teilweise ausbalanciertes, unvollständiges Blockdesign umgesetzt (Frey, Hartig & Rupp, 2009; Gonzalez & Tutkowski, 2009). Dabei treten alle Aufgaben an allen Blockpositionen mit nahezu gleicher Häufigkeit auf. Zudem sind alle Aufgaben innerhalb eines Kompetenzbereichs (Lese- bzw. Hörverstehen) direkt oder indirekt miteinander verlinkt.

Analysemethode

Um zu entscheiden, ob die Items im Mittel einfacher oder schwieriger sind, wenn die Testbearbeitung im PPA- oder im TBA-Modus erfolgte, wurden die dichotomen Itemantworten (richtig/falsch) mithilfe eines linear-logistischen Testmodells (LLTM; vgl. De Boeck, 2008; Fischer, 1973; Janssen, Schepers & Peres, 2004; Wilson & De Boeck, 2004) ausgewertet. Im LLTM wird die empirische Schwierigkeit (einer Vielzahl) von Items durch (eine geringere Menge von) Itemeigenschaften linear vorhergesagt. Die hier betrachtete Itemeigenschaft ist der Bearbeitungsmodus, der in zwei Ausprägungen vorliegt (PPA oder TBA). Im Modell bildet der Bearbeitungsmodus PPA die Referenzkategorie und der Effekt des TBA-Modus wird in Relation dazu linear modelliert. Es wird dabei jeweils angegeben, inwieweit die Bearbeitung im TBA-Modus von der Bearbeitung im PPA-Modus abweicht. Positive Werte geben an, dass die Bearbeitung im TBA-Modus leichter ausfällt und negative Werte geben an, dass die Bearbeitung im TBA-Modus schwieriger ausfällt als im PPA-Modus. Entsprechend den Empfehlungen von De Boeck et al. (2011) wurde das LLTM als ein allgemeines lineares gemischtes Modell (Generalized Linear Mixed Model [GLMM]; vgl. Molenberghs & Verbeke, 2004; Wilson & De Boeck, 2004) mit Personen und Items als Zufallsfaktoren spezifiziert.[8]

[8] Die Modellierung von Items als zufällige Faktoren erlaubt es, dass Items in ihrer Schwierigkeit variieren können, selbst wenn sie dieselbe Eigenschaft (hier: denselben Modus) haben. Die Modellierung von Personen als zufälliger Faktor erlaubt es zu berücksichtigen, dass Personen in ihrer Fähigkeit variieren können, selbst wenn sie dieselbe(n) Eigenschaft(en) (z. B. Geschlechts- oder Schulartzugehörigkeit) haben.

1.2.3 Ergebnisse

Im Folgenden wird zunächst berichtet, inwieweit die im Papierformat und im computerbasierten Format bearbeiteten Tests äquivalent sind und das gleiche Konstrukt erfassen (Konstruktäquivalenz). Anschließend wird der Frage nachgegangen, ob im Fach Englisch in den Kompetenzbereichen Lese- und Hörverstehen Moduseffekte auftreten und inwieweit diese konstant oder differenziell ausfallen. Sofern Interaktionseffekte zwischen dem Modus und dem Darstellungsformat der Items (einseitig, mehrseitig) bzw. zentralen Hintergrundmerkmalen der Schülerinnen und Schüler (Schulartbesuch, Geschlecht, Sprachhintergrund, sonderpädagogischer Förderbedarf) auftreten, liegen differenzielle Moduseffekte vor.

Konstruktäquivalenz

Betrachtet man die durch den Test gemessenen Kompetenzen im Papierformat und im computerbasierten Format als separate Konstrukte, lässt sich überprüfen, ob diese substanziell äquivalent sind – also ob Lese- bzw. Hörverstehen in gleicher Weise gemessen werden, unabhängig davon, in welchem Modus (PPA oder TBA) der Test bearbeitet wird. Dies ist dann der Fall, wenn die Zusammenhänge zwischen den Testwerten möglichst perfekt ausfallen und sich die Unterscheidung zwischen Testpersonen bzw. die Rangreihe der Testpersonen zwischen den Modi nicht unterscheidet.
Die messfehlerbereinigten (latenten) Korrelationen zwischen den Kompetenzwerten im PPA- bzw. TBA-Format betragen für Leseverstehen $r = .81$ und für Hörverstehen $r = .78$. Die durch beide Modi gemessenen Kompetenzen sind demnach zwar stark miteinander assoziiert, weichen aber signifikant von einer perfekten Korrelation von 1.00 ab und sind damit nicht äquivalent. Somit erreichen Schülerinnen und Schüler, die bei der Bearbeitung des Tests im Papierformat am besten waren, nicht in allen Fällen auch die besten Testwerte im computerbasierten Test und umgekehrt. Die nicht perfekten Korrelationen weisen auch darauf hin, dass Unterschiede in den erreichten Testwerten teilweise durch konstruktfremde Einflüsse bedingt werden. Inwieweit diese auf unterschiedliche Darstellungsformen der Items (einseitig oder mehrseitig) bzw. Merkmale der Testpersonen zurückzuführen sind, wird im Folgenden untersucht.

Moduseffekte

a) Einfluss des Darstellungsformats der Items auf Moduseffekte

Die Ergebnisse in Tabelle 1.1 weisen sowohl für das Leseverstehen als auch für das Hörverstehen auf Moduseffekte hin, die für beide Kompetenzbereiche signifikant negativ sind: Im Vergleich zur papierbasierten Testbearbeitung sind Lese- und Hörverstehensitems schwieriger, wenn sie im TBA-Format bearbeitet werden. Dieser Effekt ist jeweils größer, wenn die Items im TBA-Format auf mehreren Bildschirmseiten angeordnet sind (mehrseitige Darstellungsform), als wenn sie auf einer Bildschirmseite dargestellt werden (einseitige Darstellungsform). Deutlich wird auch, dass Moduseffekte im Leseverstehen stärker ausgeprägt sind als im Hörverstehen. So fällt der Moduseffekt im Hörverstehen bei einseitiger Darstellung zwar signifikant aus, er beträgt aber nur

-0.06 Logits. Bezogen auf die Bildungsstandardmetrik fallen die Kompetenzwerte im Hörverstehen bei einer Testbearbeitung im TBA-Format also etwa 6 Punkte niedriger aus als bei einer Testbearbeitung im Papierformat. Somit handelt es sich um einen kleinen Effekt.

Im Leseverstehen fällt der Moduseffekt sowohl bei einer einseitigen Darstellung der Items (-0.22 Logits) als auch bei einer mehrseitigen Darstellung (-0.42 Logits) deutlich stärker aus. Bezogen auf die Bildungsstandardmetrik liegen die Kompetenzwerte im Leseverstehen um etwa 22 Punkte (bei einseitiger Darstellung der Items) bzw. 42 Punkte (bei mehrseitiger Darstellung der Items) unter den Kompetenzwerten, die bei einer Testbearbeitung im Papierformat erreicht werden. Bei einem geschätzten Lernzuwachs von etwa 40 Punkten im Fach Englisch am Ende der Sekundarstufe I (Stanat, Böhme, Schipolowski & Haag, 2016), entspricht dies – je nach Darstellungsform der Items – zwischen einem halben und einem Schuljahr. Dass sich die Stärke der Moduseffekte zwischen den Kompetenzbereichen erheblich unterscheidet, spiegelt sich auch im Anteil der durch Moduseffekte aufgeklärten Varianz wider. Dieser Anteil ist im Leseverstehen dreimal so hoch wie im Hörverstehen.

Tabelle 1.1. Univariate Moduseffekte im Lese- und Hörverstehen

Parameter	Leseverstehen			Hörverstehen		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	0.98	0.11	‹.01	0.58	0.12	‹.01
TBA: einseitig	-0.22	0.02	‹.01	-0.06	0.02	‹.01
TBA: mehrseitig	-0.42	0.04	‹.01	-0.28	0.04	‹.01
Gütemaße						
Anzahl Personen	2644			2667		
marginales R^2	0.36 %			0.12 %		

Anmerkungen. Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items.

Zusammenfassend lässt sich festhalten, dass Moduseffekte nicht über Kompetenzbereiche eines Faches oder Darstellungsformate der Items (einseitig, mehrseitig) hinweg generalisiert werden können. Inwieweit Moduseffekte nicht nur durch die Darstellungsform der Items beeinflusst werden, sondern darüber hinaus von weiteren personenbezogenen Hintergrundmerkmalen, wird im Folgenden betrachtet.

b) Abhängigkeit der Moduseffekte vom Schulartbesuch

Tabelle 1.2 zeigt zunächst, dass Schülerinnen und Schüler am Gymnasium erwartungsgemäß sowohl im Lese- als auch im Hörverstehen deutlich höhere Kompetenzwerte erzielen als Jugendliche an nichtgymnasialen Schularten (1.55 bzw. 1.56 Logits). Bezogen auf die Bildungsstandardmetrik entspricht dies einem Kompetenzvorsprung von ungefähr 160 Punkten zugunsten der Gymnasiastinnen und Gymnasiasten. Darüber hinaus fallen die Haupteffekte für den Modus unter Kontrolle der Schulartunterschiede im Leseverstehen nahezu unverändert und im Hörverstehen etwas geringer aus als

zuvor. Bei einseitiger Darstellung der Items ist der Moduseffekt im Hörverstehen nicht mehr signifikant.

Ferner zeigt sich für das Leseverstehen ein signifikanter Interaktionseffekt. Demnach fällt der Moduseffekt für Gymnasiastinnen und Gymnasiasten bei einseitiger Darstellung der TBA-Items signifikant größer aus als für Jugendliche an nichtgymnasialen Schularten (-0.10 Logits). Das heißt, Jugendlichen am Gymnasium fällt der Moduswechsel schwerer als Jugendlichen an nichtgymnasialen Schularten. Aus diesem Grund reduziert sich der Kompetenzrückstand von Jugendlichen nichtgymnasialer Schularten gegenüber Jugendlichen am Gymnasium durch die computerbasierte Testbearbeitung um knapp 10 Punkte auf der Bildungsstandardmetrik. Bei einer mehrseitigen Darstellungsform zeigt sich im Leseverstehen hingegen kein Interaktionseffekt mit der Schulart.

Für das Hörverstehen sind für beide Darstellungsformen signifikante Interaktionseffekte festzustellen. Hier verringert sich der Leistungsvorsprung der Gymnasiastinnen und Gymnasiasten bei einer computerbasierten Testung gegenüber Jugendlichen an nichtgymnasialen Schularten um etwa 18 Punkte (einseitige Darstellung) bzw. 23 Punkte (mehrseitige Darstellung) im Vergleich zu einer Testbearbeitung im Papierformat. Insgesamt können die Moduseffekte somit nicht über Schularten hinweg generalisiert werden.

Tabelle 1.2. Interaktion von Moduseffekten und Schulart im Lese- und Hörverstehen

Parameter	Leseverstehen			Hörverstehen		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	0.44	0.11	‹.01	0.04	0.12	.74
Gymnasium	1.55	0.05	‹.01	1.58	0.05	‹.01
TBA: einseitig	-0.19	0.03	‹.01	-0.02	0.03	.49
TBA: mehrseitig	-0.41	0.04	‹.01	-0.22	0.04	‹.01
Interaktionseffekte						
TBA: e x Gymnasium	-0.10	0.05	‹.05	-0.18	0.04	‹.01
TBA: m x Gymnasium	0.00	0.07	.99	-0.23	0.06	‹.01
Gütemaße						
Anzahl Personen	2644			2667		
marginales R^2	8.30 %			7.00 %		

Anmerkungen. Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe Schulart = nichtgymnasiale Schulart.

c) Abhängigkeit der Moduseffekte von der Geschlechtszugehörigkeit

Inwieweit Moduseffekte für Mädchen und Jungen unterschiedlich ausfallen, ist in Tabelle 1.3 abgebildet. Zunächst zeigt sich sowohl für das Lese- als auch für das Hörverstehen ein signifikant negativer Geschlechtereffekt. Dies bedeutet, dass Jungen erwartungsgemäß in beiden Kompetenzbereichen geringere Kompetenzwerte erreichen als Mädchen, wobei der Leistungsunterschied mit knapp

36 Punkten (-0.36 Logits) im Leseverstehen etwa doppelt so groß ausfällt wie im Hörverstehen. Hier beträgt der Unterschied bezogen auf die Bildungsstandardmetrik etwa 19 Punkte (-0.19 Logits).

Unter Kontrolle der Geschlechtszugehörigkeit fallen die Haupteffekte für den Modus wie zuvor (vgl. Tabelle 1.1 und 1.2) für Lese- und Hörverstehen signifikant negativ aus, wobei sie im Hörverstehen weiterhin deutlich geringer ausgeprägt sind. Darüber hinaus treten keine statistisch bedeutsamen Interaktionseffekte auf. Richtung und Stärke von Moduseffekten unterscheiden sich zwischen Mädchen und Jungen also nicht, sondern können über beide Geschlechtergruppen hinweg generalisiert werden.

Tabelle 1.3. Interaktion von Moduseffekten und Geschlecht im Lese- und Hörverstehen

Parameter	Leseverstehen			Hörverstehen		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	1.16	0.11	<.01	0.68	0.12	< .01
männlich	-0.36	0.06	<.01	-0.19	0.06	< .01
TBA: einseitig	-0.24	0.03	<.01	-0.09	0.03	< .01
TBA: mehrseitig	-0.45	0.05	<.01	-0.31	0.05	< .01
Interaktionseffekte						
TBA: e x männlich	0.05	0.04	.21	0.05	0.04	.15
TBA: m x männlich	0.05	0.06	.41	0.06	0.06	.34
Gütemaße						
Anzahl Personen	2641			2664		
marginales R^2	0.76 %			0.21 %		

Anmerkungen. Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe Geschlecht = weiblich.

d) Abhängigkeit der Moduseffekte vom Sprachhintergrund

Tabelle 1.4 zeigt Moduseffekte in Abhängigkeit des Sprachhintergrunds der Schülerinnen und Schüler. Der negative Haupteffekt für Jugendliche, die zu Hause vorwiegend eine andere Sprache als Deutsch sprechen, entspricht unabhängig vom Modus einem Leistungsrückstand von ungefähr 69 Punkten im Leseverstehen (-0.69 Logits) bzw. 72 Punkten im Hörverstehen (-0.72 Logits) auf der Bildungsstandardmetrik im Vergleich zu Jugendlichen, die zu Hause immer oder meistens deutsch sprechen.

Unter Kontrolle des Sprachhintergrunds fallen die Haupteffekte für den Modus ähnlich wie zuvor signifikant negativ aus. Die Interaktionseffekte sind bis auf eine Ausnahme statistisch signifikant und positiv ausgeprägt. Demnach fällt die computerbasierte Testbearbeitung Jugendlichen mit nichtdeutschem Sprachhintergrund im Durchschnitt leichter als Jugendlichen mit deutschem Sprachhintergrund. Eine Ausnahme bildet lediglich die computerbasierte Testbearbeitung mit mehrseitiger

Darstellung der Items im Kompetenzbereich Leseverstehen. In allen anderen Fällen verringert sich der Leistungsrückstand zwischen Jugendlichen mit nicht-deutschem Sprachhintergrund gegenüber Jugendlichen mit deutschem Sprachhintergrund im Vergleich zu einer Testbearbeitung im Papierformat im Leseverstehen um etwa 25 Punkte (einseitige Darstellung: 0.25 Logits) bzw. im Hörverstehen zwischen 17 und 24 Punkten (einseitig: 0.17 Logits, mehrseitig: 0.24 Logits).

Auch über Jugendliche verschiedener Herkunftssprachen hinweg lassen sich Moduseffekte somit nicht generalisieren.

Tabelle 1.4. Interaktion von Moduseffekten und Sprachhintergrund im Lese- und Hörverstehen

Parameter	Leseverstehen			Hörverstehen		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	1.12	0.11	‹.01	0.71	0.12	‹.01
nicht-deutsch	-0.69	0.08	‹.01	-0.72	0.08	‹.01
TBA: einseitig	-0.27	0.03	‹.01	-0.09	0.02	‹.01
TBA: mehrseitig	-0.45	0.04	‹.01	-0.34	0.04	‹.01
Interaktionseffekte						
TBA: e x nicht-deutsch	0.25	0.06	‹.01	0.17	0.05	‹.01
TBA: m x nicht-deutsch	0.07	0.09	.41	0.24	0.08	‹.01
Gütemaße						
Anzahl Personen	2578			2581		
marginales R^2	1.08 %			0.87 %		

Anmerkungen. Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe Sprachhintergrund = deutsch.

e) Abhängigkeit der Moduseffekte vom sonderpädagogischen Förderbedarf

In Tabelle 1.5 ist dargestellt, inwieweit sich Moduseffekte für Jugendliche mit und ohne sonderpädagogischen Förderbedarf (SPF) an allgemeinbildenden Schulen unterscheiden. Zunächst zeigt sich erwartungskonform, dass Jugendliche mit SPF in beiden Kompetenzbereichen unabhängig vom Modus deutlich geringere Kompetenzwerte erreichen als Jugendliche ohne SPF. Dabei ist der Effekt im Leseverstehen mit etwa 168 Punkten (-1.68 Logits) auf der Bildungsstandardmetrik etwas stärker ausgeprägt als im Hörverstehen (145 Punkte, -1.45 Logits). Unter Kontrolle des SPF sind die Haupteffekte für den Modus nahezu unverändert signifikant negativ und im Hörverstehen geringer ausgeprägt als im Leseverstehen.

Ein signifikanter Interaktionseffekt tritt nur im Leseverstehen für die einseitige Darstellung der Items auf. Demnach ist der Moduseffekt für Schülerinnen und Schüler mit SPF geringer ausgeprägt als für Schülerinnen und Schüler ohne SPF. Jugendliche mit SPF verringern ihren Kompetenzrückstand gegenüber Jugendlichen ohne SPF im Vergleich zu einer papierbasierten Testbearbeitung bei einer

computerbasierten Testbearbeitung mit einseitiger Darstellung der Items um etwa 31 Punkte (0.31 Logits) auf der Bildungsstandardmetrik. Ähnliche Effektstärken zeigen sich tendenziell auch bei mehrseitiger Itemdarstellung im Lese- und Hörverstehen, wobei diese nicht statistisch signifikant ausfallen.

Moduseffekte können also zumindest im Leseverstehen nicht über Jugendliche mit und ohne SPF hinweg generalisiert werden.

Tabelle 1.5. Interaktion von Moduseffekten und sonderpädagogischem Förderbedarf (SPF) im Lese- und Hörverstehen

Parameter	Leseverstehen			Hörverstehen		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	1.02	0.11	<.01	0.61	0.12	<.01
SPF	-1.68	0.21	<.01	-1.45	0.20	<.01
TBA: einseitig	-0.23	0.02	<.01	-0.06	0.02	<.01
TBA: mehrseitig	-0.45	0.04	<.01	-0.29	0.04	<.01
Interaktionseffekte						
TBA: e x SPF	0.31	0.15	< .05	-0.11	0.15	.43
TBA: m x SPF	0.28	0.23	.23	0.40	0.26	.12
Gütemaße						
Anzahl Personen	2595			2618		
marginales R^2	1.02 %			0.63 %		

Anmerkungen. Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe SPF = Schülerinnen und Schüler ohne sonderpädagogischen Förderbedarf.

1.2.4 Konsequenzen

Die Prüfung der Konstruktäquivalenz weist für das Fach Englisch lediglich auf moderat hohe Zusammenhänge zwischen den papier- und computerbasierten Testformen zur Erfassung des Lese- und Hörverstehens hin, die signifikant von einer perfekten Korrelation abweichen. Somit wird weder das Lese- noch das Hörverstehen im Fach Englisch im Papierformat und im computerbasierten Format in äquivalenter Weise gemessen.

In Übereinstimmung mit internationalen und nationalen Studien (Fishbein et al., 2018; Goldhammer et al., 2019; Jerrim et al., 2018; Nachtigall, 2018) weisen auch die Ergebnisse der vorliegenden Studie auf Moduseffekte hin. Sowohl im Lese- als auch im Hörverstehen fallen die Kompetenzwerte bei einer computerbasierten Testbearbeitung durchschnittlich geringer aus als bei einer papierbasierten Testbearbeitung. Die Stärke der Moduseffekte variiert dabei je nach Darstellungsform der Items. Bei einer mehrseitigen Itemdarstellung sind die Moduseffekte überwiegend stärker ausgeprägt. Dies könnte auf Schwierigkeiten bei der Navigation hindeuten und spricht dafür, zukünftig eher die ein-

seitige (und möglicherweise intuitivere) Darstellungsform umzusetzen, bei der die Schülerinnen und Schüler zwischen Items eines Stimulus vertikal scrollen müssen.

Ein weiteres Ergebnis der vorliegenden Studie ist, dass zwar die Richtung, nicht aber die Stärke von Moduseffekten über beide Kompetenzbereiche im Fach Englisch hinweg konsistent ausfällt. Je nach Darstellungsform der Items (einseitig, mehrseitig) liegen die erreichten Kompetenzwerte im Leseverstehen zwischen 22 und 42 Punkten und im Hörverstehen zwischen 6 und 28 Punkten unter denen, die bei einer Testbearbeitung im Papierformat erreicht werden.

Darüber hinaus wurde deutlich, dass Moduseffekte zwischen Mädchen und Jungen vergleichbar ausfallen. Für Schülerinnen und Schüler unterschiedlicher Schularten, mit deutschem und nichtdeutschem Sprachhintergrund und mit bzw. ohne sonderpädagogischen Förderbedarf hingegen zeigen sich bedeutsame Interaktionseffekte. Demnach fällt insbesondere Jugendlichen am Gymnasium (vgl. auch Nachtigall, 2018), Jugendlichen mit deutschem Sprachhintergrund und Jugendlichen ohne sonderpädagogischem Förderbedarf die computerbasierte Testbearbeitung schwerer als der jeweiligen Vergleichsgruppe. Da diese Schülergruppen bei einer Testbearbeitung im Papierformat einen signifikanten und teils deutlichen Leistungsvorsprung gegenüber der jeweiligen Vergleichsgruppe aufweisen, verringert sich dieser bei einer computerbasierten Testbearbeitung, wobei die Effekte im Leseverstehen überwiegend etwas stärker ausgeprägt sind als im Hörverstehen. Inwieweit diese Effekte mit Unterschieden im Nutzungsverhalten digitaler Medien und technischer Geräte oder der Testteilnahmemotivation zusammenhängen, wird im Jahr 2020 genauer untersucht.

Insgesamt weisen die Ergebnisse also auf differenzielle Moduseffekte hin (vgl. Abschnitt 1.2.1), die nicht über Darstellungsformen der Items (einseitig, mehrseitig) und zentrale Hintergrundmerkmale der Schülerschaft (Schulart, Sprachhintergrund, SPF) hinweg generalisiert werden können. Aus diesem Grund ist es nicht möglich, eine einheitliche Verrechnungsvorschrift zu bestimmen, anhand derer Item- und Personenparameter konsistent über Darstellungsformen und Teilpopulationen hinweg adjustiert und in Bezug zu den bereits vorliegenden Kompetenzstufenmodellen gesetzt werden können. Eine Umstellung auf TBA erfordert somit eine Neunormierung der Bildungsstandards und eine Überarbeitung der Kompetenzstufenmodelle.
Nachfolgend werden Empfehlungen beschrieben, wie die Länder mit Moduseffekten umgehen können, solange nur papierbasierte Itemkennwerte vorliegen, VERA in den Ländern aber bereits computerbasiert umgesetzt wird und Ergebnisse mit Bezug zu den vorliegenden Bildungsstandards interpretiert werden.

1.2.5 Empfehlungen für den Umgang mit Moduseffekten in VERA

Ist keine Konstruktäquivalenz zwischen papier- und computerbasierten Tests gegeben, können Kompetenzwerte des einen Tests (hier TBA) nicht auf die Metrik des Referenztests (hier PPA) umgerechnet werden. Verrechnungsvorschriften für die jeweiligen Fächer und Kompetenzbereiche zu bestimmen, wäre zwar im Zuge jeweils separat durchzuführender Moduseffektstudien mit erheblichem Aufwand möglich, aber die Verlässlichkeit dieser Verrechnungsvorschriften wäre aufgrund der teils

substanziellen Interaktionseffekte so stark eingeschränkt, dass sich der dadurch entstehende Aufwand kaum durch den (ggf. nur kurz- bis mittelfristigen) Nutzen rechtfertigen ließe.

Für die Zeit bis zum Vorliegen von TBA-basierten Itemparametern schlägt das IQB vor, bei einer computerbasierten Durchführung von VERA vorläufig eine Verlinkung über Personen (anstatt Items) vorzunehmen. Eine Voraussetzung dafür ist, dass der Test in dem jeweiligen Land sowohl in der PPA- als auch in der TBA-Variante angeboten und durchgeführt wird. Beide Gruppen, die den Test entweder im PPA-Format oder im TBA-Format bearbeiten, sollten sich möglichst nicht systematisch in ihren Hintergrundmerkmalen voneinander unterscheiden. Der papierbasierte Test wird dann wie gehabt mittels der vom IQB bereitgestellten papierbasierten Itemkennwerte auf der Metrik der Bildungsstandards verankert. Anschließend wird der TBA-Test indirekt mit dem (bereits verankerten) papierbasierten Test verlinkt, wobei die (nicht überprüfbare) Annahme zu treffen wäre, dass die Kompetenzmittelwerte in beiden Teilpopulationen gleich sind. Sämtliche TBA-Itemparameter werden dann frei geschätzt und sind unbeeinflusst von (interagierenden) Moduseffekten. Zu berücksichtigen ist, dass dies nicht der Fall ist, wenn beruhend auf den TBA-Ergebnissen spezifische Ergebnisse für bestimmte Teilpopulationen ermittelt werden sollen, für die Interaktionseffekte für den Modus festgestellt wurden, wie beispielsweise für die Schulart. Aus den genannten Gründen ist die Interpretierbarkeit der Ergebnisse also auch bei diesem Vorgehen eigeschränkt und sollte deshalb nur im Übergangszeitraum angewendet werden.

1.3 Innovative Aufgabenformate

Ziel von innovativen Aufgabenformaten ist es, Anforderungen zu operationalisieren, die zwar in den Bildungsstandards beschrieben werden, bislang jedoch anhand des papierbasierten Testformats nicht oder nur sehr eingeschränkt getestet werden können (z. B. Leitidee „Funktionaler Zusammenhang“: „Veränderungen von Größen mittels Funktionen, auch unter Verwendung eines Tabellenkalkulationsprogramms beschreiben“; vgl. KMK, 2004). Die Ergebnisse der Bedarfsanalyse zur Weiterentwicklung der Bildungsstandards weisen in allen Fächern auf die Bedeutung der digitalen Bildung für den Aufbau inhaltsbezogener Kompetenzen hin. Dies betrifft etwa den Umgang mit digitalen Hilfsmitteln wie Onlinequellen und Onlinewörterbüchern in den Fremdsprachen oder die Anwendung von Symbolsprache und mathematischen Werkzeugen im Fach Mathematik. Um daraus abzuleitende Anforderungen valide in Testaufgaben abbilden zu können, dürfte es bei einer entsprechenden Weiterentwicklung der Bildungsstandards zukünftig notwendig sein, in allen Fächern neben den herkömmlichen Aufgabenformaten auch innovative Testaufgaben sowie illustrierende Lernaufgaben im TBA-Format zu entwickeln.

Erste Erfahrungen mit der Entwicklung und Erprobung innovativer Aufgabenformate sammelt das IQB in der aktuellen Projektlaufzeit im Fach Mathematik. Um eine valide Abdeckung der Anforderungen in den Bildungsstandards zu erzielen, wurde analog zur papierbasierten Aufgabenentwicklung in VERA eine Aufgabenentwicklungsgruppe gebildet, der vier Personen (mehrheitlich Lehrkräfte), ein externer Bewerter (Prof. Dr. Siller, Universität Würzburg) und der fachdidaktische Kooperationspartner des Arbeitsbereichs VERA-8 Mathematik (Prof. Dr. Greefrath, Universität Münster) angehören. In einem Zeitraum von 12 Monaten wurden 31 innovative Testaufgaben mit insgesamt 95 Items

entwickelt, wobei zunächst primär die Leitideen „Messen“ und „Raum und Form“ im Fokus standen. Das am IQB umgesetzte Vorgehen bei der Entwicklung innovativer Aufgabenformate orientiert sich sowohl am regulären Aufgabenentwicklungsprozess im Rahmen von VERA als auch am Vorgehen internationaler Studien wie eTIMSS (Cotter, 2019).

Entsprechend des Antrags zur Sicherstellung der Anschlussfähigkeit des TBA-Projekts werden im Jahr 2020 Ergebnisse darüber vorliegen, unter welchen Bedingungen Aufgabenentwürfe von den Aufgabenentwicklerinnen und Aufgabenentwicklern direkt im Online-Autorenmodul des IQB-Testsystems erstellt, hier von den fachdidaktischen Kooperationspartnern kommentiert und anschließend überarbeitet werden können und welche (Unterstützungs-) Strukturen dafür ggf. geschaffen werden müssen (z. B. ob Arbeitstreffen der Aufgabenentwicklerinnen und -entwickler am IQB teilweise durch digitale Austauschformen bzw. Feedbacktools ersetzt werden können). Außerdem wird der Frage nachgegangen, inwieweit mit innovativen Mathematikaufgaben das gleiche Kompetenzkonstrukt erfasst wird wie mit regulären Mathematikaufgaben.

Die Erfahrungen aus der aktuellen Projektlaufzeit weisen auf folgende Aspekte hin, die zukünftig bei der Entwicklung innovativer Aufgabenformate berücksichtigt werden sollten:

- Innovative Aufgabenformate sollten immer dann entwickelt werden, wenn reguläre Aufgabenformate nicht ausreichend sind, um die Bildungsstandards hinreichend inhaltsvalide abzubilden. Wie hoch der Anteil innovativer Aufgabenformate an der Gesamtzahl zu entwickelnder Aufgaben zukünftig sein wird, hängt von unterschiedlichen Faktoren ab, wie beispielsweise den zu operationalisierenden Anforderungen im Bereich digitaler Bildung. In eTIMSS weist etwa ein Drittel der eingesetzten Items im Fach Mathematik ein innovatives Format auf (Cotter, 2019) und in der Machbarkeitsstudie wurde der Bedarf an innovativen Aufgabenformaten bereits auf Basis der aktuell vorliegenden Bildungsstandards für das Fach Mathematik von dem fachdidaktischen Kooperationspartner auf ca. 25 % geschätzt, um die Kompetenzkonstrukte in der Sekundarstufe I hinreichend inhaltsvalide in Testaufgaben abzubilden. Folglich wäre für den mathematisch-naturwissenschaftlichen Bereich von einer Größenordnung zwischen 25 % und 33 % auszugehen.
- Um innovative Aufgabenformate zu entwickeln, müssen sowohl die fachdidaktischen Kooperationspartner als auch mindestens einzelne Personen in der Aufgabenentwicklergruppe des jeweiligen Fachs über Expertise bzw. Erfahrungen im Bereich der digitalen Bildung verfügen, worauf bei der Zusammenstellung der Aufgabenentwicklergruppen zukünftig zu achten wäre.
- Die Erstellung von innovativen Aufgabenformaten wird derzeit in Kooperation mit dem DIPF unter Nutzung des Testsystems „ItemBuilder“ erprobt (vgl. Abschnitt 1.1.1). Die Erfahrungen aus der Aufgabenentwicklung zeigen, dass die Erstellung der Aufgaben in diesem System perspektivisch deutlich zu aufwendig wäre, da dies momentan nur durch sehr gut geschultes Personal mit Programmierkenntnissen möglich ist. Aktuell übertragen diese Personen die Papierentwürfe der Aufgabenentwicklerinnen und -entwickler in das TBA-Format. Dadurch entstehen zusätzliche, zeitintensive Rückmeldeschleifen, die in einem typischen VERA-Aufgabenentwicklungsprozess nicht realisiert werden könnten. Für eine effiziente Entwicklung innovativer Aufgabenformate in

eng getakteten Zeitfenstern wie in VERA und den Bildungstrends muss der Entwicklungsprozess weiter optimiert werden. Ergebnisse zur Frage, wie dies realisiert werden kann, werden Ende 2020 vorliegen. Zudem sollte das deutlich niedrigschwelliger zu bedienende Autorenmodul des IQB-Testsystems sukzessive um die Funktionsbereiche erweitert werden, die perspektivisch voraussichtlich häufiger bei der Entwicklung innovativer Aufgabenformate benötigt werden (z. B in Mathematik: Lineal, Taschenrechner, dynamische Geometriesoftware). Andernfalls ist dauerhaft mit aufwendigen Ergänzungsprogrammierungen zu rechnen oder es muss vollständig auf die Entwicklung innovativer Aufgabenformate verzichtet werden. Letzteres ist insbesondere vor dem Hintergrund der Bedeutung digitaler Bildung, die bei der Weiterentwicklung der Bildungsstandards eine Rolle spielen werden, nicht zu empfehlen.

- Die neu entwickelten innovativen Mathematikaufgaben wurden im November 2019 in insgesamt 11 Klassen aus neun Schulen (N = 229) in Berlin, Brandenburg und Nordrhein-Westfalen einer ersten technischen Erprobung unterzogen. Dabei wurden unter Verwendung des ItemBuilder (DIPF), der erstmals online eingesetzt wurde, bereits beim Systemcheck teilweise technische Schwierigkeiten beim Laden komplexer integrierter Anwendungen (OnlyOffice, GeoGebra) festgestellt. Zudem kam es während der Testdurchführung wiederholt zu Browserabstürzen. Auch dies spricht dafür, den Funktionsumfang des IQB-Testsystems aufgrund der bisher stabileren Leistungsfähigkeit um Anwendungen zu erweitern, die zukünftig vorrausichtlich häufiger benötigt werden.
- Trotz der angestrebten niedrigschwelligen Bedienbarkeit des IQB-Testsystems ist davon auszugehen, dass Aufgabenentwicklerinnen und -entwickler zukünftig im Umgang mit dem Autorenmodul geschult werden müssen, um einen Einblick in die Konstruktion innovativer Aufgabenformate und den Umgang mit unterschiedlichen Gestaltungselementen zu gewinnen. Diese Schulungen werden zukünftig regelmäßig in den VERA-Aufgabenentwicklungsprozessen erforderlich sein, da wie bisher von einer Fluktuation der beteiligten Personen auszugehen ist.

1.4 Fazit

Zusammenfassend weisen die Ergebnisse der Machbarkeitsstudie darauf hin, dass die Durchführung von onlinebasierten Kompetenztests unter Nutzung von Schulhardware mit dem IQB-Testsystem zuverlässig möglich ist. Das Testsystem hat sich bewährt und Grundlagen für die Anschlussfähigkeit an ländereigene Systeme wurden aufgebaut. Da anzunehmen ist, dass sich die Ausstattung mit Endgeräten in Schulen in den kommenden Jahren u. a. durch Mittel des Digitalpakts weiter verbessern wird, kann eine wesentliche Grundvoraussetzung für die Umstellung des Testmodus am IQB auf TBA und somit die Anschlussfähigkeit an nationale und internationale Standards als gegeben angesehen werden.

Die Ergebnisse der vorliegenden Studie zeigen aber auch, dass Lese- und Hörverstehen im Fach Englisch im papier- und computerbasierten Format nicht in gleicher Weise erfasst werden können und Moduseffekte auftreten, die weder über Kompetenzbereiche innerhalb eines Fachs noch über zentra-

le Hintergrundmerkmale der Schülerschaft (Schulartbesuch, Sprachhintergrund, SPF) und Darstellungsformate der Items hinweg generalisiert werden können. Da in der vorliegenden Studie differenzielle Moduseffekte beobachtet wurden, kann anhand der Ergebnisse nicht unmittelbar auf Richtung und Stärke von Moduseffekten in anderen Kompetenzbereichen geschlossen werden. Für die Umstellung von VERA auf TBA bedeutet dies, dass Moduseffekte für alle Fächer und Kompetenzbereiche in ressourcenintensiven Moduseffektstudien separat untersucht und adjustiert werden müssten, wenn die Ergebnisse mit Bezug zu den vorliegenden Bildungsstandards interpretiert werden sollen. Da Kosten und Nutzen bei diesem Vorgehen aus Sicht des IQB nicht gerechtfertigt sind, sollte im Rahmen von VERA vorläufig eine Verlinkung über Personen statt über Items vorgenommen werden bis TBA-basierte Itemkennwerte vorliegen. Die Umstellung auf TBA sollte zeitlich und inhaltlich gekoppelt an die Weiterentwicklung der Bildungsstandards durchgeführt werden, um doppelte Arbeitsprozesse zu vermeiden (wiederholte Normierungsstudien). Ein weiterer Grund hierfür ist, dass die Bedarfsanalyse zur Weiterentwicklung der Bildungsstandards für alle Fächer auf die Bedeutung von Kompetenzen im Bereich der digitalen Bildung hinweist. Es ist anzunehmen, dass die überarbeiteten Bildungsstandards nur dann hinreichend inhaltsvalide in Testaufgaben und in illustrierenden Lernaufgaben abgebildet werden können, wenn dazu auch innovative TBA-Aufgabenformate verwendet werden (z. B. Simulationen, Recherchetools, Onlinewörterbücher).

Literatur

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B. & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment, 6*.

Chen, G., Cheng, W., Chang, T.-W., Zheng, X. & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*, 213-225.

Cotter, K. (2019). *Evaluating the validity of the eTIMSS 2019 mathematics problem solving and inquiry tasks.* Boston College, Boston.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-539.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Francis, T. & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of statistical software, 39*, 1-28.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fishbein, B., Martin, M. O., Mullis, I. V. S. & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*, 11.

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

Goldhammer, F., Harrison, S., Bürger, S., Kroehne, U., Lüdtke, O., Robitzsch, A., Köller, O., Heine, J.-H. & Mang, J. (2019). Vertiefende Analysen zur Umstellung des Modus von Papier auf Computer. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018. Schülerleistungen im internationalen Vergleich* (S. 163-186). Münster: Waxmann.

Gonzalez, E. & Tutkowski, L. (2009). 125 principles of multiple matrix booklet designs and parameter recovery in large-scale assessments *Educational Measurement: Issues and Practice, 28*, 39-53.

Hassler Hallstedt, M. & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment, 23*, 195-210.

Hofer, S., Holzberger, D., Heine, J.-H., Reinhold, F., Schiepe-Tiska, A., Weis, M. & Reiss, K. (2019). Schulische Lerngelegenheiten zur Sprach- und Leseförderung im Kontext der Digitalisierung. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018: Grundbildung im internationalen Vergleich* (S. 111-128). Münster: Waxmann.

Janssen, R., Schepers, J. & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 189-212). New York, NY: Springer New York.

Jerrim, J., Micklewright, J., Heine, J.-H., Sälzer, C. & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Journal of Education, 44*, 476–493.

Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 22*, 22-37.

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14*, 169.

Margolin, S., Driscoll, C., Toland, M. & Kegler, J. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology, 27*.

Molenberghs, G. & Verbeke, G. (2004). An introduction to generalized (non)linear mixed models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 111-153). New York: Springer.

Nachtigall, C. (2018). *Thüringer Kompetenztest. Landesbericht Schuljahr 2017/18*. Jena: Projekt kompetenztest.de.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F. & Heine, J.-H. (2016). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten. *Diagnostica, 63*, 148-165.

Sandene, B., Bennett, R. E., Braswell, J. & Oranje, A. (2005). Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project. Washington DC: U.S. Department of Education, National Center for Education Statistics.

Schulz-Heidorf, K. & Støle, H. (2018). Gender differences in Norwegian PIRLS 2016 and ePIRLS 2016 results at test mode, text and item format level. *Nordic Journal of Literacy Research, 4*, 167-183.

Singer, L. M. & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*, 155-172.

Stanat, P., Böhme, K., Schipolowski, S. & Haag, N. (Hrsg.). (2016). *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich*. Münster: Waxmann.

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-24.

Weis, M., Zehner, F., Strohmaier, A., Artelt, C. & Pfost, M. (2016). Lesekompetenz in PISA 2015: Ergebnisse, Veränderungen und Perspektiven. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015. Eine Studie zwischen Kontinuität und Innovation* (S. 249-283). Münster: Waxmann.

Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (S. 43-74). New York: Springer.

Teil B: Ergebnisse der TBA-Machbarkeitsstudie 2020

Um die Anschlussfähigkeit des TBA-Projekts sicherzustellen wurde die Projektlaufzeit bis zum 31.12.2020 verlängert. Ziel war es, das Online-Testsystem technisch zu optimieren und weiter zu erproben, um die Stabilität, Leistungsfähigkeit und Nutzungsfreundlichkeit zu erhöhen. Zudem sollten die vorliegenden Daten aus der TBA-Machbarkeitsstudie 2018-2019 vertiefend ausgewertet und dokumentiert werden. Nachfolgend werden die zentralen Ergebnisse aus dem Verlängerungszeitraum beschrieben. Diese fließen in den kommenden Jahren in die Umstellung der Testinstrumente des IQB bzw. die weitere Optimierung des Testsystems ein.

1. Weiterentwicklung und Dokumentation des IQB-Testsystems

Das Online-Testsystem des IQB (www.iqb-testcenter.de) wurde im Jahr 2019 für die Durchführung der Machbarkeitsstudie eingesetzt und verfügte deshalb zunächst nur über die dafür erforderlichen Funktionen. Um die Stabilität, Leistungsfähigkeit und den Funktionsumfang für einen langfristigen Einsatz des Testsystems am IQB und in den Ländern sicherzustellen, wurden im Jahr 2020 die zentralen Webanwendungen zur Aufgabenentwicklung und -bearbeitung (Teststudio), Testdurchführung (Testcenter) und Datenaufbereitung (Dateneditor) weiterentwickelt und erprobt. Aktuell steht eine erprobte Basisversion des Testsystems zur Verfügung. Für VERA 2021 wurden den Ländern erstmals zusätzlich zu den papierbasierten Aufgaben auch TBA-Aufgaben für das Fach Englisch übergeben, die unter Nutzung des IQB-Testsystems bzw. einzelner Webanwendungen oder eines eigenen Systems in Kombination mit der länderübergreifend standardisierten Verona-Schnittstelle eingesetzt werden können. Konkret wurden die folgenden technischen Weiterentwicklungen umgesetzt:

Aufgabenentwicklung

Die Stabilität und Nutzungsfreundlichkeit der Programmierungen zur Web-Anwendung „IQB-Teststudio“, die für die Aufgabenentwicklung und -bearbeitung genutzt werden, wurden im Jahr 2020 verbessert. Dies betraf vor allem die Optimierung und Nutzungsfreundlichkeit der Installation. Hierzu wurde eine neue Anwendung programmiert, mit deren Hilfe das IQB-Teststudio unkompliziert auf einem neuen (z. B. ländereigenen) Server installiert werden kann. Die Installation umfasst jetzt auch die automatische Einrichtung einer Datenbank. Die Einbindung von digitalen Hilfsmitteln (z. B. Taschenrechner) und die Einführung eines einheitlichen Metadatenmanagements wurden konzeptionell weiter vorangetrieben und die Verona-Schnittstelle um eine Editorkomponente zur Anpassung und Bearbeitung von bereitgestellten Aufgaben erweitert. Darüber hinaus wurden umfassende Schulungsmaterialien (Videotutorials, Manuale) für Aufgabenentwickler*innen erarbeitet und auf dem öffentlich zugänglichen Codeverwaltungsportal GitHub bereitgestellt.

Testdurchführung

Die Programmierungen zur Web-Anwendung „IQB-Testcenter“, das zur Testvorbereitung und -durchführung eingesetzt wird, betrafen im Jahr 2020 umfassende Verbesserungen der Codequalität. Insbesondere die serverseitige Anwendungsschnittstelle (API) wurde umfassend überarbeitet und mit

einer standardisierten Dokumentation versehen, die das automatische Testen der Codequalität erlaubt (z. B. Aufdecken von Datenflussanomalien).

Außerdem wurde eine Testleitungskonsole entwickelt, die neben einem Systemcheck für die Testvorbereitung einen Gruppenmonitor umfasst, dessen Ansicht sich regelmäßig automatisch aktualisiert und die Lehrkraft bzw. Testleitung dabei unterstützt, den Bearbeitungsfortschritt aller teilnehmenden Schüler*innen zu beobachten und zu steuern.[9] Darüber hinaus wurden folgende Funktionen neu implementiert bzw. optimiert:

- Der Systemcheck wurde grundlegend überarbeitet. Die Eignung eines Systems, das für den Test verwendet werden soll (z. B. die digitale Ausstattung in der Schule), kann jetzt zuverlässiger abgeschätzt werden. Der darauf basierende technische Bericht an das IQB wurde nutzungsfreundlicher gestaltet.
- Es wurde eine Setup-Anwendung entwickelt, durch die die Installation des Testcenters automatisiert und an länderspezifische Bedürfnisse angepasst werden kann (z. B. Nutzung landesspezifischer Logos, Begrüßungstexte und weiterer Textmeldungen während des Testverlaufs).
- Das Testcenter bemerkt ein zeitweises Verlassen des Tests durch die Testperson (z. B. Surfen im Internet). Dieses Verlassen wird in den Logdaten dokumentiert. Außerdem wird die Testleitung über die Testleitungskonsole darüber informiert und kann die Testbearbeitung sperren bzw. die Testperson darauf ansprechen.
- Es wurde ein Demomodus implementiert, mit dem ausgelieferte Tests des IQB von Dritten (z. B. den VERA-Verantwortlichen in den Ländern) begutachtet und getestet werden können. Antworten und Logdaten werden nicht gespeichert. Der Demomodus wurde mit der Übergabe der Englisch-Testhefte für VERA 2021 erstmals erprobt. Perspektivisch sollen darüber auch Rückmeldungen der Länder zur Aufgabenqualität eingeholt werden (u. a. in Vorbereitung auf das jährliche Treffen „Diskussion der Pilotierungsaufgaben“).
- Sämtliche Programmierungen zum Testcenter wurden umfassend dokumentiert und deren Funktionalität und Einsatz in Videotutorials erläutert, die auf GitHub zugänglich sind.

Datenaufbereitung

Für die Datenaufbereitung wurde die Software „ITC-Toolbox“ programmiert und bereitgestellt. Der Konvertierungsprozess der unterschiedlichen Datenformate, die im Rahmen der Testung erhoben werden und in der Machbarkeitsstudie noch sehr aufwendig manuell aufbereitet werden mussten, erfolgt nun automatisiert. Dabei wird ein Primärdatensatz erzeugt, der mit allen gängigen Statistikprogrammen (z. B. R, SPSS) ohne weitere Zwischenschritte weiterverarbeitet werden kann und neben den Rohdaten der Testaufgaben (z. B. gewählte Antwortoption eines Multiple-Choice Items)

[9] Im Projektzeitraum 2021 – 2023 werden typische Szenarien der Testdurchführung beschrieben und datenschutzrechtlich bewertet. Bei der Planung und Durchführung von Studien wie z. B. VERA muss das jeweilige Vorgehen dokumentiert werden. Bei einzelnen Studien kann es erforderlich sein, die Zustimmung der Schüler*innen oder von Erziehungsberechtigten einzuholen, wenn die Testleitungskonsole verwendet wird.

auch Prozessdaten (z. B. Bearbeitungszeiten) enthält, die aus den Logdaten ermittelt und dem Primärdatensatz als Variablen hinzugefügt werden.

Technologische Infrastruktur

Das IQB hat im Jahr 2020 die Nutzung externer Cloud-Server ausgebaut. Die Stabilität und Performance sind überzeugend, deshalb wird das IQB diese Form des Hostings auch zukünftig nutzen. Die Installation erfolgt jetzt automatisiert (über Docker auf virtuelle Server), wodurch das manuelle Übertragen der Programmierungen auf dateibasierte Server entfällt.

Die Dokumentation auf dem Codeverwaltungssystem GitHub wurde umfassend erweitert. Auf dieser Plattform sind die jeweils aktuellen Programmierungen des IQB öffentlich zugänglich[10]. Es werden automatische Qualitätstests des bereitgestellten Codes durchgeführt, technische Schwierigkeiten werden über Tickets (sog. Issuetracker über die Plattform ZenHub) schnell bearbeitet. Auch die Dokumentation (Wiki, Manuale, Videotutorials) des Testsystems wurde hier hinterlegt und wird kontinuierlich ausgebaut. Der Metadatenkatalog soll ebenfalls auf GitHub bereitgestellt werden.

Für die Qualitätssicherung der technischen Programmierungen des IQB wurden im Berichtszeitraum mehrere Aufträge an einen externen Berater vergeben. Es erfolgten Begutachtungen mit Empfehlungen zur Optimierung von Verona-Player (Bundling, Schnittstelle), Verona-Editor (Schnittstelle, Userinterface), Modularisierung, Single-Sign-On und Metadatensystem. Diese Empfehlungen wurden bzw. werden vom IQB implementiert.

Im Verlängerungszeitraum wurden aufgrund der pandemiebedingten Einschränkungen nur zwei der drei geplanten Workshops mit Vertreter*innen der Länder bzw. der Fachgruppe VERA Online durchgeführt, die im Onlineformat stattfanden. Der Schwerpunkt der Workshops lag auf den Themen Installation der IQB-Anwendungen, Weiterentwicklung der länderübergreifenden Verona-Schnittstelle und Funktionsbereiche zur Testdurchführung. Die Präsentationen des IQB wurden vorab als Video produziert und stehen auch weiterhin öffentlich auf GitHub zur Verfügung.

Unter Leitung des IQB hat die Fachgruppe VERA Online auf Bitte der VERA-Steuergruppe im Jahr 2020 ein Eckpunktepapier erarbeitet, aus dem die bereits getroffenen Vereinbarungen sowie noch zu klärende Fragen zur technischen Umsetzung hervorgehen. Das Eckpunktepapier macht deutlich, dass die Fachgruppe VERA Online die technischen Leitlinien des IQB-Testsystems unterstützt.

Der Austausch mit anderen nationalen und internationalen Expert*innen auf dem Gebiet des technologiebasierten Testens musste im Jahr 2020 pandemiebedingt eingeschränkt werden, wobei die Zusammenarbeit mit dem TBA-Zentrum des DIPF ausgebaut wurde. Aktuell finden monatliche Treffen der Software-Entwickler*innen statt. Zur technischen Validierung (sog. *proof-of-concept*) der länderübergreifenden Verona-Schnittstelle hat das DIPF eine eigene Testeinheit des DIPF-ItemBuilder mit der Verona-Schnittstelle versehen. Diese konnte anschließend erfolgreich in das IQB-Testcenter geladen und dort angezeigt und bearbeitet werden. Ein Austausch unterschiedlicher Itemformate,

[10] https://github.com/iqb-berlin

die über die gemeinsame Schnittstelle verfügen, ist somit möglich. Darüber hinaus hat das IQB in einer Kooperation mit der Universität Tübingen (Prof. Dr. Detmar Meurer, Dr. Ramon Ziai) die automatische Kodierung von offenen Kurzantworten anhand der Daten aus der IQB-Machbarkeitsstudie erprobt. Die Ergebnisse verschiedener Kodierungsansätze (u. a. *Bag of Words*, *String Similarities*), die auf dem Konzept des Natural Language Processing basieren, waren ausgesprochen vielversprechend. Die Übereinstimmung mit den Ergebnissen aus einer (deutlich aufwendigeren) händischen Kodierung variierte zwischen 93,80 % und 96,80 %. Die Kooperation mit der Eberhard Karls Universität Tübingen wird fortgeführt, um die technische Anbindung sowie die Kodierung komplexerer Freitextantworten weiter zu erproben.

2. Erprobungsstudien

Im Jahr 2020 wurden zwei Studien durchgeführt, in denen die Stabilität und Praxistauglichkeit des IQB-Testsystems unter alltagspraktischen Bedingungen weiter erprobt sowie innovative Mathematikitems entwickelt und getestet wurden.

2.1 Erprobung der Testleitungskonsole

Der Einsatz des IQB-Testsystems und die Durchführungsmodalitäten (Systemcheck zur Testvorbereitung und Durchführung des TBA-Tests) sollten im Jahr 2020 erstmals durch Lehrkräfte erprobt werden. Dazu war geplant, im Rahmen der regulären VERA-Pilotierung 2020 eine zusätzliche Ergänzungsstichprobe von etwa 20 Schulen zu rekrutieren, die im Fach Englisch computerbasierte Testaufgaben bearbeiten. Durch den pandemiebedingten Ausfall der regulären VERA-Pilotierung musste die Erhebung verschoben werden.

Die Testung fand im Herbst 2020 mit ca. 472 Schüler*innen in 20 achten Klassen an 16 Schulen in Baden-Württemberg, Brandenburg und Hamburg statt. In der 45-minütigen Testzeit bearbeiteten die Schüler*innen TBA-Aufgaben im Leseverstehen und Hörverstehen im Fach Englisch. Zusätzlich erfolgte eine Online-Befragung der Lehrkräfte, die den Test mit dem Systemcheck vorbereiteten und ihn mit der Testleitungskonsole durchführten.

Bis auf wenige technische Probleme verlief die Erprobung erfolgreich. Die Stabilität und Funktionalität der Testleitungskonsole und des IQB-Testsystems konnten damit bestätigt werden.

Zur Vorbereitung der Testsitzung wurden den Schulen und Lehrkräften verschiedene Materialien zur Verfügung gestellt. Die Anleitung für den Systemcheck wurde weitgehend als hilfreich bewertet und die Lehrkräfte gaben überwiegend an, sich gut auf den Test vorbereitet gefühlt zu haben. Insbesondere die bereitgestellte Videopräsentation zur Vorbereitung auf die Testdurchführung wurde, wenn sie angesehen wurde, als sehr hilfreich bewertet. Die Steuerung der Testdurchführung mit Hilfe der Testleitungskonsole gelang überwiegend gut, Schwierigkeiten hatten Lehrkräfte zum Teil beim Freischalten eines neuen Aufgabenblocks (z. B. beim Wechsel vom Hör- zum Leseverstehen). Vergleichsweise häufig wurde aus diesem Grund auch die während der Testdurchführung bereitgestellte technische Hotline angerufen, die sich bewährt hat. Die Nutzungsfreundlichkeit der Testleitungs-

konsole wird auf Grundlage der Rückmeldungen der Lehrkräfte, die die Testungen durchgeführt haben, weiterentwickelt und verbessert. Geplant ist unter anderem, den Anmeldeprozess und das Freischalten von Aufgabenblöcken zu vereinfachen.

Die teilnehmenden Schulen erhielten eine Rückmeldung über die von den Schüler*innen erzielten Leistungen im computerbasierten Englischtest und über die Ergebnisse der Lehrkräftebefragung.

2.2 Innovative Mathematikaufgaben

Erprobung des Aufgabenentwicklungsprozesses

Im Rahmen dieses Teilprojektes wurde zunächst der Frage nachgegangen, unter welchen Bedingungen innovative TBA-Aufgabenentwürfe durch Aufgabenentwickler*innen digital in einem lokal installierten Aufgabeneditor erstellt werden können. Ziel war es, aus den Erfahrungen dieses vollständig digitalen Entwicklungsprozesses Schlüsse für die zukünftige Organisation des Aufgabenentwicklungsprozesses (z. B. Eignung von Austauschplattformen, Onlinetagungen, Feedbacktools, benötigte Unterstützungsstrukturen) und die Gestaltung des Aufgabeneditors (insbesondere, welche Funktionen und Werkzeuge benötigt werden) abzuleiten.

An der Aufgabenentwicklung beteiligten sich insgesamt vier in der Aufgabenentwicklung erfahrene Mathematiklehrkräfte, der fachdidaktische Kooperationspartner für VERA-8 Mathematik (Prof. Gilbert Greefrath, Universität Münster), sowie zwei wissenschaftliche Mitarbeiter*innen. Im Projektzeitraum wurden 27 neue, innovative Aufgaben entwickelt, mit denen Anforderungen aus dem aktuell vorliegenden Kompetenzstufenmodell der Bildungsstandards für den Mittleren Schulabschluss im Fach Mathematik operationalisiert werden sollten, die bisher gar nicht oder nicht hinreichend im papierbasierten Format erfasst werden können. Im Mittelpunkt standen Aufgaben zu den Leitideen Raum und Form, Funktionaler Zusammenhang sowie Daten und Zufall. Da das IQB-Teststudio aktuell noch nicht über einen Aufgabeneditor verfügt, der die benötigten digitalen Hilfsmittel (insbesondere GeoGebra, Excel) bereitstellt, wurde dieses Teilprojekt in Kooperation mit dem DIPF unter Nutzung des DIPF-ItemBuilder durchgeführt. Für die Arbeit mit dem ItemBuilder wurden die Aufgabenentwickler*innen und der fachdidaktische Kooperationspartner umfassend geschult, während der Projektphase eng begleitet (u. a. im Rahmen von vier Aufgabenentwicklungstagungen) und zu ihren Erfahrungen im Umgang mit dem ItemBuilder und zum Verlauf des Aufgabenentwicklungsprozesses befragt. Auf diese Weise sollte auch ermittelt werden, inwieweit der ItemBuilder zukünftig für die Entwicklung innovativer Items in die Aufgabenentwicklung des IQB eingebunden werden könnte.

Der Entwicklungsprozess verlief insgesamt erfolgreich, alle Aufgaben konnten planmäßig bereitgestellt werden. Aus der Begleitung und Befragung der Teilnehmenden können folgende Schlussfolgerungen für die zukünftige Aufgabenentwicklung am IQB abgeleitet werden:

- Die Bedienung des DIPF-ItemBuilders ist für den regulären Aufgabenentwicklungsprozess am IQB aktuell nicht ausreichend nutzungsfreundlich für durchschnittlich bis wenig technikaffine Lehrkräfte. Die Teilnehmenden bewerteten die Arbeit mit dem ItemBuilder trotz umfassender Schulung und Betreuung während der technischen Umsetzung als kompliziert, unverständlich,

verwirrend und eher schwer zu lernen. Der Mehraufwand für einen Aufgabenentwurf wurde mit durchschnittlich 180 Minuten angegeben. Der ItemBuilder ist deshalb derzeit als nicht geeignet für die Aufgabenentwicklung am IQB einzuschätzen. Da eine nutzungsfreundlichere Weiterentwicklung des ItemBuilder am DIPF derzeit nicht geplant ist, wird das IQB zentrale, fächerübergreifend relevante digitale Hilfsmittel, die einen mittleren Interaktionsgrad erfordern (z. B. *Drag-and-Drop*, Tabellenkalkulation, dynamische Geometrie-Software, Taschenrechner, Markierfunktionen) und voraussichtlich häufiger für die Aufgabenentwicklung und -bearbeitung der weiterentwickelten Bildungsstandards benötigt werden, im Aufgabeneditor des IQB-Teststudios umsetzen. Die Bedienung des Teststudios wird dabei auch weiterhin auf die Bedürfnisse von Lehrkräften abgestimmt und intuitiv, ohne spezielle Vorkenntnisse zu bedienen sein. Lediglich für innovative Aufgabenformate mit einem sehr hohen Interaktionsgrad, die voraussichtlich nur punktuell für die Operationalisierung sehr komplexer Anforderungen benötigt werden (voraussichtlich ca. 5-10 % der Aufgaben), wird das IQB auf den ItemBuilder des DIPF zurückgreifen. Der Aufgabenentwicklungsprozess erfolgt dann abweichend vom regulär geplanten Vorgehen (Entwicklung durch Aufgabenentwickler*innen im Testsystem), indem die technische Umsetzung eines Aufgabenentwurfs durch die EDV unterstützt wird.

- Als verbesserungswürdig wurde zudem die Nutzung verschiedener digitaler Austausch- und Kommunikationswerkzeuge betrachtet, weil der Entwicklungsprozess dadurch bislang insgesamt zu unübersichtlich ist (z. B. GitLab zur Organisation des Aufgabenentwicklungsprozesses einschließlich technischer Support und zur Kommunikation und Kommentierung der Aufgaben, Word-Dokumente zur Dokumentation der Metadaten inklusive Musterlösungen, ItemBuilder zur Aufgabenbearbeitung, Browser zum Erzeugen des Aufgaben-Previews). Das IQB wird deshalb sukzessive alle Funktionen in das Teststudio integrieren, die während des Aufgabenentwicklungsprozesses von unterschiedlichen Akteuren benötigt werden (z. B. Preview einer Aufgabe, Integration von Musterlösungen, ein-/ausblendbare Metadaten einschließlich des aktuellen Bearbeitungsstatus, Möglichkeiten zur Kommentierung während der Bewertungs- und Begutachtungsphase, Funktionsprüfung).
- Von den vier Aufgabenentwicklungstagungen fanden drei planmäßig als Videokonferenz statt. Es zeigte sich, dass diese eine gute Ergänzung von Präsenzveranstaltungen für einen kurzfristigen Austausch darstellen, wenn sie nicht länger als zwei bis drei Stunden dauern und ggf. auf mehrere Tage verteilt werden. Digitale Austauschformate können Präsenzveranstaltungen aber insbesondere bei der Entwicklung von Aufgabenideen und kreativen Überarbeitungen nicht ersetzen. Deshalb werden auch zukünftig Präsenztreffen notwendig sein.

Erprobung innovativer Mathematikaufgaben

Eine Auswahl der 27 neu entwickelten innovativen Mathematikaufgaben sollte gemeinsam mit regulären Mathematikaufgaben in einer Studie im Herbst 2020 erprobt werden. Da innovative Aufgabenformate aufgrund der Einbettung digitaler Anwendungen höhere technische Anforderungen an das verwendete System stellen als reguläre Aufgaben (z. B. aufgrund größerer Datenmengen), sollten technische Aspekte mit der vorhandenen Schulhardware getestet werden. Ein weiteres Ziel bestand

darin, die Güte der innovativen Aufgaben genauer zu untersuchen. Unter anderem sollte der Frage nachgegangen werden, ob beide Aufgabentypen auf einer gemeinsamen Skala verortet werden können und das gleiche inhaltliche Konstrukt abbilden.

Die Erprobungsstudie wurde im Herbst 2020 durchgeführt. Aufgrund des pandemiebedingten Unterrichtsausfalls musste die Testzeit von zwei Stunden auf eine Stunde reduziert werden. Zur Entlastung der Lehrkräfte wurden ausschließlich externe Testleitungen mit der Durchführung der Tests beauftragt. Insgesamt nahmen an der Erprobung 15 achte Klassen mit 352 Schüler*innen aus Baden-Württemberg, Bayern, Brandenburg und Nordrhein-Westfalen teil. Aufgrund der verringerten Testzeit war es nicht möglich, sowohl innovative als auch reguläre Mathematikaufgaben einzusetzen, sodass die strukturelle Validität der Aufgaben nicht untersucht werden konnte. Da die Ergebnisse der ersten Erprobung im Jahr 2019 die Annahme nahelegten, dass Schüler*innen Schwierigkeiten damit haben, die Werkzeuge der eingebetteten digitalen Anwendungen (insbesondere Excel, GeoGebra) gezielt zu verwenden (vgl. Frenken et al., 2020), wurden die innovativen Aufgaben erneut erprobt. Ziel war es, neben den technischen Aspekten auch das Bearbeitungsverhalten genauer zu untersuchen. Der Fokus lag dabei primär auf Aufgaben, bei denen mit dynamischer Geometriesoftware (z. B. Veränderung eines Flächeninhalts) und Tabellenkalkulation (z. B. Erstellung eines Diagramms) gearbeitet werden musste. Die Analyse dieser Daten erfolgt in Zusammenarbeit mit den Kooperationspartner*innen an der Universität Münster und kann aufgrund von derzeit abzuschließenden Qualifikationsarbeiten erst im Sommer 2021 beendet werden. Die Befragung der Lehrkräfte der teilnehmenden Klassen lässt jedoch die vorsichtige Annahme zu, dass Schüler*innen im Mathematikunterricht bisher selbst kaum eigene Erfahrungen im Umgang mit diesen Anwendungen sammeln und deshalb vermutlich wenig vertraut mit deren Handhabung sind. Zwar gaben die Lehrkräfte an, punktuell sowohl Excel- als auch GeoGebra-Anwendungen für die Veranschaulichung von mathematischen Zusammenhängen im Unterricht zu nutzen, die Bearbeitung von derartigen Aufgaben durch Schüler*innen findet aber nicht oder kaum statt. Auch aus diesem Grund wird es wichtig sein, im Rahmen der Weiterentwicklung der Bildungsstandards Lernaufgaben zu entwickeln, die exemplarisch veranschaulichen, wie Lehrkräfte fachspezifisch relevante digitale Kompetenzen im Unterricht aufbauen können

2.3 VERA-Arbeitstagung

Es war geplant, die Erprobungen des Testsystems in den Ländern durch einen Austausch über Erfahrungen und Anforderungen zu intensivieren, die bei Fortführung des Umstellungsprozesses der Testinstrumente des IQB berücksichtigt werden müssen. Für einen breiten Austausch mit unterschiedlichen Akteursgruppen (Länder, auswertende Einrichtungen, Aufgabenentwickler*innen, Fachkoordinator*innen des IQB) sollte u. a. die VERA-Arbeitstagung im Jahr 2020 genutzt werden, die vom 22. bis 23.06.2020 am IQB stattfinden sollte. Aufgrund der pandemiebedingten Einschränkungen wurde die Durchführung der Tagung in der 418. Sitzung des Schulausschusses am 26.03.2020 abgesagt. Die Tagung wird am 14. und 15. Juni 2021 mit dem gleichen Themenschwerpunkt als Onlineveranstaltung nachgeholt.

3. Vertiefte Auswertung methodischer und inhaltlicher Fragestellungen

Im Verlängerungszeitraum wurden die vorliegenden Daten aus der TBA-Machbarkeitsstudie vertiefend ausgewertet. Dabei wurde zum einen untersucht, inwieweit die festgestellten Moduseffekte beim Wechsel vom papier- zum computerbasierten Testen anhand der selbsteingeschätzten Fähigkeiten im Umgang mit digitalen Medien und technischen Geräten erklärt werden können (Abschnitt 3.1). Zum anderen wurden Prozessdaten ausgewertet, die zusätzliche diagnostische Informationen darüber liefern können, ob während der Testbearbeitung Schwierigkeiten auftraten und inwieweit diese mit Itemeigenschaften bzw. Personenmerkmalen zusammenhängen (Abschnitt 3.2).

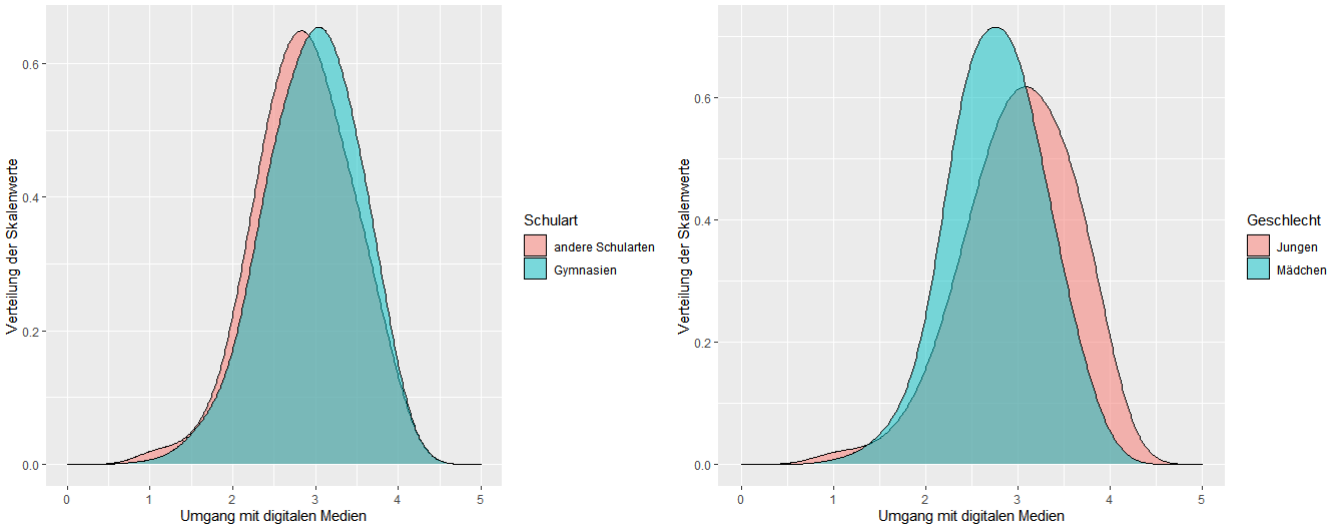
3.1 Erklärung von Moduseffekten durch selbsteingeschätzte Fähigkeiten im Umgang mit digitalen Medien und technischen Geräten

Moduseffekte zwischen papier- und computerbasierten Testergebnissen können mit unterschiedlichen individuellen (z. B. Vertrautheit im Umgang mit der Maus oder Tastatur bzw. digitalen Hilfsmitteln) und administrationsspezifischen Merkmalen (Darstellung der Items) zusammenhängen (Clariana & Wallace, 2002; Kröhne & Martens, 2011). In der TBA-Machbarkeitsstudie wurde u. a. deutlich, dass die Stärke der Moduseffekte in Abhängigkeit vom Darstellungsformat der Items einer Aufgabe (einseitig: Scrollen vs. mehrseitig: Blättern) und der besuchten Schulart (Gymnasium vs. nicht-gymnasiale Schulen) variierte. So zeigte sich, dass sowohl im Lese- als auch im Hörverstehen im Fach Englisch starke Moduseffekte auftraten, wenn die Items eines Stimulus mehrseitig dargestellt wurden und die Schüler*innen zur Bearbeitung des nächsten Items jeweils auf eine neue Seite weiterklicken mussten. Im Vergleich dazu fiel den Schüler*innen die Bearbeitung der TBA-Aufgaben leichter, wenn alle Items eines Stimulus bzw. einer Aufgabe auf einer Seite abgebildet waren und die Schüler*innen zwischen Items vertikal scrollen mussten. Ebenso wurden u. a. stärkere Moduseffekte für Gymnasiast*innen im Vergleich zu Jugendlichen aus nicht-gymnasialen Schulen erkennbar. Im Jahr 2020 wurde untersucht, inwieweit diese Moduseffekte durch die Fähigkeitsselbsteinschätzung der Schüler*innen im Umgang mit digitalen Medien und technischen Geräten erklärt werden können.

Dazu gaben die Schüler*innen für insgesamt 11 Items auf einer vierstufigen Skala (1 = *„sehr unsicher“* bis 4 = *„sehr sicher“,* α = .83) an, wie sicher sie ihrer Einschätzung nach im Umgang mit digitalen Medien und technischen Geräten sind. Dabei wurden sowohl basale (z. B. eine Mail mit Anhang versenden, eine Präsentation erstellen) als auch eher fortgeschrittene Fähigkeiten bewertet (einen Newsfeed abonnieren, Programmieren).

Der empirische Mittelwert der Skala lag bei 2.90 und damit über dem theoretischen Skalenmittelwert von 2.50. Die Jugendlichen schätzen ihre Fähigkeiten im Umgang mit digitalen Medien und technischen Geräten also insgesamt als eher sicher ein. Abbildung 1 veranschaulicht die empirische Verteilung der Skalenwerte für Jugendliche unterschiedlicher Schularten bzw. für Mädchen und Jungen. Während keine bedeutsamen Unterschiede in der Fähigkeitsselbsteinschätzung zwischen Schüler*innen am Gymnasium (M = 2.96, SD = 0.54) und an nicht-gymnasialen Schulen (M = 2.86, SD = 0.57; d = 0.17) erkennbar sind, schätzen Jungen (M = 3.01, SD = 0.59) ihre Kompetenzen signifikant höher ein als Mädchen (M = 2.78, SD = 0.50; d = 0.43).

Abbildung 1. Verteilung der Fähigkeitsselbsteinschätzung im Umgang mit digitalen Medien und technischen Geräten nach Schulart und Geschlecht.



Bezogen auf die Erklärung der Moduseffekte zeigt sich für den Bereich *Hörverstehen* im Fach Englisch unter Berücksichtigung der selbsteingeschätzten Fähigkeit im Umgang mit digitalen Medien und technischen Geräten zunächst ein Haupteffekt für dieses Merkmal (vgl. Tabelle 1, Modell 2). Das heißt, Schüler*innen mit einer höheren Fähigkeitsselbsteinschätzung erreichen auch bessere Ergebnisse im Hörverstehen. Der Moduseffekt für die mehrseitige Darstellung der Items verringert sich unter Kontrolle der Fähigkeitsselbsteinschätzung nur geringfügig und bleibt weiterhin statistisch signifikant (vgl. Tabelle 1, Effekte in Modell 1 vs. Modell 2). Dieser Moduseffekt scheint also mit weiteren, hier nicht betrachteten Merkmalen zusammenzuhängen. Ebenso zeigen die Interaktionseffekte zwischen Darstellungsformat und Schulart, dass Gymnasiast*innen auch unter Kontrolle der Fähigkeitsselbsteinschätzung stärker von Moduseffekten im Hörverstehen betroffen sind als Jugendliche nicht-gymnasialer Schulen. Gymnasiast*innen fällt der Wechsel auf TBA im Hörverstehen also weiterhin schwerer als Jugendlichen in nicht-gymnasialen Schulen. Interaktionseffekte zwischen der Darstellungsform der Items und der Fähigkeitsselbsteinschätzung treten nicht auf. Insgesamt trägt die Kontrolle der Fähigkeitsselbsteinschätzung somit nicht substanziell zur Aufklärung der Moduseffekte im Hörverstehen bei.

Tabelle 1. Moduseffekte im Hörverstehen

Parameter	Modell 1			Modell 2		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	0.04	0.12	.74	0.02	0.14	.88
Gymnasium	1.58	0.05	‹.01	1.70	0.06	‹.01
TBA: einseitig	-0.02	0.03	.49	-0.01	0.03	.72
TBA: mehrseitig	-0.22	0.04	‹.01	-0.17	0.05	‹.01
FSE	--	--	--	0.35	0.05	‹.01
Interaktionseffekte						
TBA: e x Gymnasium	-0.18	0.04	‹.01	-0.21	0.05	‹.01
TBA: m x Gymnasium	-0.23	0.06	‹.01	-0.37	0.07	‹.01
TBA: e x FSE	--	--	--	-0.03	0.04	.51
TBA: m x FSE	--	--	--	-0.01	0.06	.90
Gütemaße						
Anzahl Personen	2667			2377		
marginales R^2	7.00 %			7.80 %		

Anmerkungen. Die Ergebnisse in Modell 1 entsprechen denen im Ergebnisbericht (Stand: 05.12.2019); Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe Schulart = nicht-gymnasiale Schulart; FSE = Fähigkeitsselbsteinschätzung im Umgang mit digitalen Medien und technischen Geräten.

Für den Bereich *Leseverstehen* im Fach Englisch zeigt sich unter Berücksichtigung der selbsteingeschätzten Fähigkeit im Umgang mit digitalen Medien und technischen Geräten ebenfalls ein Haupteffekt für dieses Merkmals (vgl. Tabelle 2). Wie im Hörverstehen erreichen Schüler*innen mit einer höheren Fähigkeitsselbsteinschätzung bessere Ergebnisse im Leseverstehen. Die Haupteffekte für die ein- und mehrseitige Darstellung der Items bleiben unter Kontrolle der Fähigkeitsselbsteinschätzung in ihrer Ausprägung nahezu unverändert. Hingegen fällt der Interaktionseffekt zwischen einseitiger Darstellung der Items und der Schulart unter Kontrolle der Fähigkeitsselbsteinschätzung nicht mehr signifikant aus.

Weiterhin tritt ein kleiner Interaktionseffekt zwischen der einseitigen Darstellung der Items und der Fähigkeitsselbsteinschätzung auf. Dieser fällt allerdings kontraintuitiv aus, weil er darauf hindeutet, dass Moduseffekte für Personen mit einer höheren Fähigkeitsselbsteinschätzung stärker ausfallen als für Personen mit geringerer Fähigkeitsselbsteinschätzung und Personen, die sich selbst als fähiger einschätzen größere Schwierigkeiten bei einem Wechsel von papier- auf computerbasierte Tests haben.

Tabelle 2. Moduseffekte im Leseverstehen

Parameter	Modell 1			Modell 2		
	Estimate	SE	*p*	Estimate	SE	*p*
Haupteffekte						
Intercept	0.44	0.11	‹.01	0.63	0.12	‹.01
Gymnasium	1.55	0.05	‹.01	1.65	0.06	‹.01
TBA: einseitig	-0.19	0.03	‹.01	-0.27	0.03	‹.01
TBA: mehrseitig	-0.41	0.04	‹.01	-0.42	0.05	‹.01
FSE				0.29	0.05	‹.01
Interaktionseffekte						
TBA: e x Gymnasium	-0.10	0.05	‹.05	0.02	0.05	.76
TBA: m x Gymnasium	0.00	0.07	.99	-0.13	0.08	.11
TBA: e x FSE				-0.09	0.04	.03
TBA: m x FSE				-0.04	0.07	.57
Gütemaße						
Anzahl Personen	2644			2369		
marginales R^2	8.30 %			9.74 %		

Anmerkungen. Die Ergebnisse in Modell 1 entsprechen denen im Ergebnisbericht (Stand: 05.12.2019); Estimate = Schätzer (Effekte in Logits); SE = Standardfehler; *p* = Signifikanzwert; TBA: einseitig/TBA: e = einseitige Darstellung der Items; TBA: mehrseitig/TBA: m = mehrseitige Darstellung der Items; Referenzgruppe Schulart = nicht-gymnasiale Schulart; FSE = Fähigkeitsselbsteinschätzung im Umgang mit digitalen Medien und technischen Geräten.

Insgesamt fallen die Moduseffekte unter Berücksichtigung der Fähigkeitsselbsteinschätzung im Umgang mit digitalen Medien und technischen Geräten für die Bereiche Lese- und Hörverstehen unterschiedlich aus. Im Hörverstehen trägt die Fähigkeitsselbsteinschätzung kaum dazu bei, den Moduseffekt der mehrseitigen Itemdarstellung aufzuklären und auch die stärkeren Moduseffekte der Gymnasiast*innen bleiben unverändert erhalten. Im Leseverstehen konnte nur der stärkere Moduseffekt der einseitigen Itemdarstellung bei Gymnasiast*innen durch die Kontrolle der Fähigkeitsselbsteinschätzung aufgeklärt werden. Dass Moduseffekte in der ein- und mehrseitigen Darstellung in Gymnasien im Hörverstehen, nicht aber im Leseverstehen, stärker ausfallen als in nicht-gymnasialen Schularten, ist auf den ersten Blick ein erwartungswidriger Befund. Er weist darauf hin, dass Gymnasiast*innen im Umgang mit individuell digital dargebotenen Hörstimuli weniger vertraut sind als Jugendliche in nicht-gymnasialen Schulen. Wie diese Unterschiede konkret zustande kommen, lässt sich anhand der vorliegenden Daten jedoch nicht bestimmen.

Insgesamt zeigen die Ergebnisse, dass die Schwierigkeiten, die beim Wechsel von papier- auf computerbasiertes Testen für alle Schüler*innen entstehen, nur geringfügig damit zusammenhängen, wie vertraut sie im Umgang mit digitalen Medien und technischen Geräten sind bzw. wie gut sie ihre Fähigkeiten in diesem Bereich beurteilen.

3.2 Prozessdaten

Ein bedeutsamer Vorteil computerbasierter Testungen besteht darin, dass neben den Antwortdaten der Schüler*innen (z. B. richtig/falsch) auch sogenannte Prozessdaten automatisch während der Bearbeitung in Logfiles gespeichert und in die Datenanalyse einbezogen werden können. Prozessdaten umfassen beispielsweise Bearbeitungszeiten von Aufgaben bzw. einzelner Items einer Aufgabe sowie Interaktionen zwischen Testperson und Testumgebung (z. B. Mausbewegungen, Klickverhalten, Texteingaben), die Rückschlüsse auf das Bearbeitungsverhalten und auf Lösungsstrategien erlauben, aber auch dazu beitragen können, die Messpräzision der Kompetenzschätzung zu erhöhen (Goldhammer & Kröhne, 2020).

Besonders kurze Bearbeitungszeiten eines Items, die nur wenige Sekunden umfassen und 10-20% unter der durchschnittlichen Bearbeitungszeit eines Items liegen, deuten beispielsweise auf sogenanntes „Low Test-Taking Engagement" bzw. „Rapid Guessing Behavior" (im Folgenden RGB) hin (Goldhammer et al., 2017; Wise & Kong, 2005). In diesem Fall geben Testteilnehmende eine Antwort auf ein dargebotenes Item in nur wenigen Sekunden nach dessen Präsentation.[11] Da es unwahrscheinlich ist, dass Schüler*innen ein Item in so kurzer Zeit gewissenhaft bearbeiten, lässt sich RGB als Ausdruck von geringer Anstrengungsbereitschaft interpretieren. Es wird angenommen, dass ein solches Bearbeitungsverhalten besonders dann auftritt, wenn eine Testperson nicht motiviert ist, engagiert an der Testung teilzunehmen (z. B. bei *Low-Stakes*-Tests wie VERA, deren Ergebnisse nicht in die Benotung einfließen) oder im Laufe der Testung ermüdet oder in Zeitdruck gerät (Kong et al., 2007). Im Rahmen der Machbarkeitsstudie wurden sowohl die Bearbeitungszeit des gesamten Tests als auch die Bearbeitungszeit für jedes einzelne Item einer Aufgabe im mehrseitigen Darstellungsformat erfasst.[12] Ziel der vertieften Analyse war es, zunächst grundlegende Qualitätsanalysen durchzuführen und zu prüfen, ob es Hinweise darauf gibt, dass Schwierigkeiten bei der Testbearbeitung bestimmter Itemformate und Schüler*innengruppen auftraten. Im weiteren Analyseprozess soll geprüft werden, ob die Messpräzision der Kompetenzschätzung durch die Berücksichtigung der Bearbeitungszeiten verbessert werden kann.

Die Ergebnisse zeigen, dass die *Bearbeitungszeit des gesamten Tests*, für den 60 Minuten vorgesehen waren, durchschnittlich bei 40.39 Minuten (*SD* = 8.91) lag. Bezogen auf einzelne Schüler*innengruppen traten zwar statistisch bedeutsame Unterschiede zwischen Mädchen (*M* = 40.75 *SD* = 8.96) und Jungen (*M* = 40.06, *SD* = 8.85, t = 2.03, $p < .05$), Schüler*innen mit SPF (*M* = 42.39, *SD* = 10.20) und ohne SPF (*M* = 40.27, *SD* = 8.86, t = 2.46, $p < .05$) sowie Jugendlichen an Gymnasi-

[11] In den vorliegenden Analysen beinhaltet dies sowohl Rapid Guessing Behavior (kurze Bearbeitungszeit und Item beantwortet) als auch Rapid Omission Behavior (kurze Bearbeitungszeit und Item nicht beantwortet), da in der Literatur von ähnlichen, zugrundeliegenden Prozessen ausgegangen wird Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1), 1-25. https://doi.org/10.1186/s40536-017-0051-9 und Auslassungsraten in VERA- bzw. Bildungstrend-Pilotierungen allgemein niedrig sind.

[12] Für das einseitige Darstellungsformat (Scrollen) war eine itemspezifische Erfassung der Bearbeitungszeit nicht möglich, da alle Items einer Aufgabe auf einer Seite abgebildet wurden und nur die Bearbeitungszeit der gesamten Aufgabe erfasst werden konnte.

en (M = 41.50, SD = 9.28) und an nicht-gymnasialen Schulen (M = 39.56, SD = 8.48, t = 5.27, p < .01) auf, diese Unterschiede bewegen sich aber lediglich im Bereich weniger Minuten und sind insofern praktisch vernachlässigbar.

Für die mehrseitige Darstellungsform (Blättern) wurden zusätzlich die *Bearbeitungszeiten der einzelnen Items* einer Aufgabe ausgewertet, da dies Anhaltspunkte darauf liefern kann, inwieweit während der Testbearbeitung eine verringerte Anstrengungsbereitschaft (RGB) zu beobachten ist und ob dieses Verhalten mit bestimmten Itemeigenschaften (z. B. offenes Antwortformat) oder Merkmalen der Schüler*innen (z. B. Schulart, Geschlecht) zusammenhängt. Die vorliegenden Analysen zeigen, dass RGB häufiger bei *Multiple-Matching*-Aufgaben (Zuordnungsaufgaben) im Kompetenzbereich Leseverstehen im Fach Englisch auftrat (im Hörverstehen wurde dieses Itemformat nicht eingesetzt). Es wäre möglich, dass diese Art der Itemgestaltung für das computerbasierte Format in der mehrseitigen Darstellungsform nicht optimal ist und die Bearbeitung deshalb häufiger mit einer geringeren Anstrengungsbereitschaft einherging als andere Itemformate (Multiple-Choice- bzw. True-False-Aufgabentypen und komplexere Aufgabenformate wie *Short-Answer-Questions* und *Table*- sowie *Sentence-Completion*). Um RGB bei Zuordnungsaufgaben künftig möglichst zu vermeiden, könnten statt der Auswahl einer Antwortalternative aus einem *Drop-Down*-Menü eher *Drag-and-Drop*-Itemformate eingesetzt werden, deren Bearbeitung interaktiver und ggf. intuitiver ist.

Die Ergebnisse zeigen zudem, dass RGB bei 11,68 % aller Antworten aufgetreten ist, die Bearbeitungszeit lag hier jeweils 15% unter der durchschnittlichen Bearbeitungszeit des jeweiligen Items. Dieses (*Normative-Threshold*-)Kriterium hat sich als zuverlässig für die Bestimmung von RGB erwiesen, da es in der Regel keine gewissenhafte Bearbeitung zulässt (Lindner et al., 2019). Dabei wurde erkennbar, dass Jungen (12.80 %) signifikant häufiger RGB zeigten als Mädchen (10.40 %, χ^2 = 48.37, df = 1, p < .01). Ebenso fiel der Anteil von RGB bei Jugendlichen nicht-gymnasialer Schulen (13.00 %) signifikant höher aus als bei Schüler*innen am Gymnasium (9.10 %, χ^2 = 113.64, df = 1, p < .01).

Zu berücksichtigen ist bei den hier dargestellten Ergebnissen, dass sich diese nur auf die mehrseitige Darstellungsform der Items beziehen, deren Bearbeitung den Schüler*innen insgesamt schwerer fiel als die Bearbeitung des TBA-Tests in der einseitigen Darstellungsform (Scrollen). Die stärkeren Moduseffekte bei der mehrseitigen Darstellungsform ließen sich weder im Hör- noch im Leseverstehen auf die Fähigkeitsselbsteinschätzung der Schüler*innen zurückführen. Da die mehrseitige Itemdarstellung mit unterschiedlichen Schwierigkeiten verbunden ist, soll zukünftig die einseitige Darstellungsform eingesetzt werden, die insgesamt weniger stark von Moduseffekten betroffen war. Eine Herausforderung stellt dabei die Erfassung von Prozessdaten dar. Um diese in der einseitigen Darstellungsform nicht nur auf Aufgabenebene zu erfassen, wie dies in der Machbarkeitsstudie umgesetzt wurde, sondern auch auf Itemebene, soll die Funktion des magnetischen Scrollens (Scroll Snap) implementiert werden. Auf diese Weise kann die Ansicht einzelner Items beim vertikalen Scrollen fixiert werden, sodass Prozessdaten so erfasst werden können, als ob Items auf jeweils separaten Seiten dargestellt sind.

Zukünftig könnte es sinnvoll sein, die VERA-Ergebnisrückmeldungen um Ergebnisse aus der Analyse von Prozessdaten zu erweitern und diese in die Ergebnisreflexion und die Ableitung von Maßnahmen

zur Unterrichtsentwicklung einzubeziehen. Anhand dieser Ergebnisse können Lehrkräfte beispielsweise erkennen, ob schwache Testergebnisse von Schüler*innen mit einer geringeren Anstrengungsbereitschaft während der Testbearbeitung einhergehen oder feststellen, welche Bearbeitungsstrategien zu Falschlösungen führten.

Um RGB entgegenzuwirken, wäre es zudem sinnvoll, VERA-Tests zukünftig stärker adaptiv zu gestalten, beispielsweise im Rahmen eines *Multi-Stage-Testdesigns* (Luecht & Sireci, 2011). In diesem Fall werden den einzelnen Schüler*innen Aufgabenmodule auf Basis der Testergebnisse aus einem vorangegangenen Aufgabenmodul adaptiv über mehrere Stufen hinweg zugewiesen. Erst durch dieses Vorgehen wird das zentrale Ziel der Modularisierung – eine optimale Passung zwischen dem individuellen Kompetenzniveau und dem Anforderungsniveau des Tests herzustellen – tatsächlich erreicht. Adaptives Testen erhöht die Testeffizienz und ist mit zahlreichen Vorteilen verbunden (u. a. Verbesserung der Messpräzision auf individueller Ebene, Vermeidung von Über- und Unterforderung, erhöhte Testteilnahmemotivation), stellt aber auch einige technische und methodische Herausforderungen an die Bereitstellung und Durchführung des Tests und erfordert neben technischen Entwicklungsarbeiten (u. a. automatische Kodierung von Freitextantworten und Scoring von Aufgabenmodulen, Routingmechanismen) auch die Durchführung von Simulationsstudien zur Beantwortung methodischer Fragestellungen (z. B. geeignetes Vorgehen zur Expositionskontrolle, um weiterhin Rückmeldungen auf Klassenebene sicherzustellen) sowie Erprobungen im Schulkontext. Da es sich dabei um eine Weiterentwicklung der technischen Infrastruktur des IQB-Testsystems handelt, sollte aus Sicht des IQB angestrebt werden, ein solches Vorhaben erneut über den DigitalPakt Schule umzusetzen.

Literatur

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602. https://doi.org/10.1111/1467-8535.00294

Frenken, L., Libbrecht, P., Greefrath, G., Schiffner, D., & Schnitzler, C. (2020). Evaluating educational standards using assessment "with" and "through" technology. In A. Donevska-Todorova, E. Faggiano, J. Trgalova, Z. Lavicza, R. Weinhandl, A. Clark-Wilson, & H. G. Weigand (Eds.), *Mathematics education in the digital age (MEDA) PROCEEDINGS*. Johannes Kepler University.

Goldhammer, F., & Kröhne, U. (2020). Computerbasiertes Assessment. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 119-141). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_6

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, *5*(1), 1-25. https://doi.org/10.1186/s40536-017-0051-9

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606-619. https://doi.org/10.1177/0013164406294779

Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*(2), 169-186. https://doi.org/10.1007/s11618-011-0185-4

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources [Original Research]. *Frontiers in Psychology*, *10*(1533), 1-15. https://doi.org/10.3389/fpsyg.2019.01533

Luecht, R. M., & Sireci, S. G. (Eds.). (2011). *A review of models for computer-based testing. Research report 2011-12*. The College Board.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2