Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2023.01.02
Volume 15 (2023), No. 1, 30–50

Armin Jentsch, Christin Beese & Knut Schwippert

Papier- oder computerbasierte Kompetenztests? Eine Generalisierbarkeitsstudie zu Moduseffekten in Deutschland im Rahmen von TIMSS 2019

Zusammenfassung

Die zunehmende Digitalisierung in Deutschland ist insbesondere in den Lernumwelten von Schüler:innen zu beobachten. So haben auch Schulvergleichsuntersuchungen zuletzt von papier- auf computerbasierte Tests umgestellt. In TIMSS 2019 wurde diese Umstellung durch eine Moduseffektstudie begleitet. An dieser Studie nahmen 2 Jahre vor der Haupterhebung 847 Viertklässler:innen teil und bearbeiteten nach dem Zufallsprinzip entweder zuerst den papier- oder den computerbasierten Kompetenztest in Mathematik und den Naturwissenschaften. Wir gehen in diesem Beitrag der Frage nach, inwieweit der Lösungserfolg bei der Aufgabenbearbeitung vom Erhebungsmodus abhängt. Wir führen dazu erstens eine Generalisierbarkeitsstudie durch und bestimmen den Anteil der Varianz in den Schüler:innenantworten, der auf Moduseffekte zurückgeht. Zweitens wird untersucht, ob die Zusammenhänge zwischen schwierigkeitsgenerierenden Merkmalen der Testaufgaben und dem Lösungserfolg der Schüler:innen von Moduseffekten betroffen sind. Die Ergebnisse zeigen, dass die Erfassung der Schüler:innenkompetenzen in beiden Domänen mit kleinen Moduseffekten einhergeht. Der Lösungserfolg ist in den computerbasierten Leistungstests etwas niedriger (Mathematik: OR = *0.87,* p = *.006; Naturwissenschaften:* OR = *0.86,* p = *.006). In Mathematik gilt dies besonders für Aufgaben in den Bereichen Daten und Problemlösen, für die Naturwissenschaften lässt sich keine entsprechende Aussage treffen. Die Bedeutsamkeit dieser Befunde für Schulvergleichsuntersuchungen und die unterrichtliche Praxis in Deutschland wird im Beitrag kritisch diskutiert.*

Dr. Armin Jentsch (Korrespondenzautor), ORCID: 0000-0002-2423-3955, University of Oslo, Department of Teacher Education and School Research, Postbox 1099 Blindern, 0317 Oslo, Norwegen
E-Mail: armin.jentsch@ils.uio.no

Christin Beese · Prof. Dr. Knut Schwippert, Universität Hamburg, Fakultät für Erziehungswissenschaft, Von-Melle-Park 8, 20146 Hamburg
E-Mail: christin.beese@uni-hamburg.de
knut.schwippert@uni-hamburg.de

Schlagworte
Generalisierbarkeitstheorie, Large-scale Assessment, Validität, Moduseffekte

Paper- or Computer-Based Achievement Tests? A Generalizability Study on Mode Effects in Germany in the Context of TIMSS 2019

Abstract
The increasing digitization in Germany can be observed particularly in students' learning environments. Large-scale assessments have also recently switched from paper-based to computer-based tests. In TIMSS 2019, this change was accompanied by a mode effect study 2 years before the main survey. In this study, 847 fourth-graders were randomly assigned to take either the paper-based or the computer-based achievement test in mathematics and science first. In this paper, we address the question to what extent student achievement depends on the survey mode. We firstly conduct a generalizability study to assess the amount of variance in item responses that is explained by mode effects. Secondly, we investigate how much the relations between difficulty-generating characteristics of the test items and student achievement differ by mode. The results show that the assessment of student achievement in both domains is affected by small mode effects. In the computer-based tests, student achievement is estimated to be slightly lower (mathematics: OR = *0.87,* p = *.006; science:* OR = *0.86,* p = *.006). In mathematics, this holds true for items addressing data and problem-solving in particular. For science, we cannot make a similar claim. In the paper, the relevance of these findings for large-scale assessments and educational practice in Germany is critically discussed.*

Keywords
generalizability theory, large-scale assessment, validity, mode effects

1. Einleitung

Die zunehmende Digitalisierung ist in Deutschland in allen Bereichen der Gesellschaft und somit auch in den Lernumgebungen von Schüler:innen zu beobachten. Vor allem die Auswirkungen der Covid-19-Pandemie haben ein Umdenken von präsenzbasiertem zu digitalem Unterricht erfordert. Im Rahmen dieses Veränderungsprozesses haben in den letzten Jahren mehrere internationale Schulvergleichsuntersuchungen von papierbasierten (*paper-based test*, PBT) zu computerbasierten Leistungstests (*computer-based test*, CBT) gewechselt: die International Computer and Information Literacy Study (ICILS) bereits in 2013, das Programme for International Student Assessment (PISA) in 2015, die Trends in Mathematics and Science Study (TIMSS) im Zyklus 2019 (Schwippert et al., 2020) und die Progress in

International Reading Literacy Study (PIRLS) im Jahr 2021 (Bos et al., 2014; Reiss, Sälzer, Schiepe-Tiska, Klieme & Köller, 2016).

Einerseits bietet die Umstellung zu CBT wichtige Vorteile, wie beispielsweise die Möglichkeit, innovative Aufgabenformate mit interaktiven Funktionen einzusetzen, sowie die Option, Schüler:innenantworten automatisiert und somit ökonomischer zu erfassen (Cotter, Centurino & Mullis, 2020; Frey & Hartig, 2013). Andererseits muss bei der Untersuchung zeitlicher Veränderungen von Schüler:innenkompetenzen auch sichergestellt werden, dass die Interpretation der Testergebnisse nicht durch die veränderten Testbedingungen gefährdet werden (American Educational Research Association [AERA], American Psychological Association & National Council on Measurement in Education, 1999). Problematisch erscheint hierbei vor allem, dass sich Eigenschaften von Testaufgaben durch Moduswechsel ändern können (z. B. Itemschwierigkeit oder -trennschärfe). Dies birgt die Gefahr, dass Unterschiede in den Testergebnissen fälschlicherweise auf inter- oder intraindividuelle Leistungsunterschiede der Schüler:innen zurückgeführt werden, in Wirklichkeit aber mit dem Erhebungsmodus zusammenhängen. *Moduseffekte* sind daher Unterschiede in Testergebnissen, die auf Erhebungsmodi zurückzuführen sind (z. B. papier- und computerbasierte Leistungstests). Kröhne und Martens (2011) gehen im Übrigen davon aus, dass es sich bei Moduseffekten stets um eine Mischung von Effekten handelt, die aus den Eigenschaften der Testaufgaben resultieren. Von den US-amerikanischen pädagogisch-psychologischen Fachgesellschaften wird die Überprüfung von Moduseffekten daher als Standardverfahren in empirischen Untersuchungen empfohlen, die von Moduswechseln betroffen sind (AERA et al., 1999; vgl. auch Kröhne & Martens, 2011).

Es liegen bereits einige empirische Befunde zu Moduseffekten bei der Erhebung von Schüler:innenkompetenzen vor, die wir nachfolgend zusammenfassen. In einer Metaanalyse finden Wang und Kolleg:innen (2007) vernachlässigbare Modusunterschiede in den durch CBT und PBT gemessenen mathematischen Kompetenzen von Schüler:innen an US-amerikanischen Schulen. Dabei berücksichtigen sie 44 Studien, von denen etwa zwei Drittel experimentell durchgeführt wurden. Ergebnisse aus Large-scale-Untersuchungen weisen dagegen darauf hin, dass die mittels PBT erhobenen Schüler:innenkompetenzen unabhängig von der Domäne gegenüber dem CBT höher ausfallen (Fishbein, Martin, Mullis & Foy, 2018; Goldhammer et al., 2019; Robitzsch et al., 2017, 2020; Zinn, Landrock & Gnambs, 2021). Testaufgaben werden gemäß diesen Befunden seltener korrekt gelöst, wenn sie im computerbasierten Format präsentiert werden. Die Trennschärfen unterscheiden sich allerdings häufig nur marginal zwischen den Erhebungsmodi und es liegen auch selten differenzielle Unterschiede in den Eigenschaften der Testaufgaben vor (für eine Ausnahme s. Goldhammer et al., 2019). Ein solches Ergebnis würde anzeigen, dass sich die Eigenschaften einzelner Testaufgaben zwischen den Erhebungsmodi unterscheiden und könnte daher die Interpretation von CBT und PBT als Erfassung einer gemeinsamen zugrundeliegenden Fähigkeit gefährden.

Robitzsch et al. (2020) konstatieren, dass die Ergebnisse zu Modusunterschieden aus Large-scale-Untersuchungen nur schwer zu deuten sind, wenn kein experimentelles Untersuchungsdesign vorliegt. Ferner fehlen häufig geeignete unabhängige Variablen, um die gefundenen Unterschiede zu erklären (Wang, Jiao, Young, Brooks & Olson, 2007). Dies nehmen wir zum Anlass für den vorliegenden Beitrag, in dem wir die genannten Desiderate in einer randomisierten Studie aufgreifen und der Frage nachgehen, inwiefern die mathematischen und naturwissenschaftlichen Testergebnisse von Schüler:innen in TIMSS 2019 vom Erhebungsmodus abhängen. Wir führen dazu eine Generalisierbarkeitsstudie (G-Studie; Cronbach, Gleser, Nanda & Rajaratnam, 1972) durch, um zu bestimmen, inwieweit die Testergebnisse von einem Moduswechsel beeinflusst sein könnten. Wir untersuchen außerdem, ob die Zusammenhänge zwischen schwierigkeitsgenerierenden Merkmalen der Testaufgaben und dem Lösungserfolg der Schüler:innen von Moduseffekten betroffen sind.

2. Forschungsanliegen

Die TIMS-Studie ist eine internationale Schulvergleichsuntersuchung, die Kompetenzen von Schüler:innen in Mathematik und Naturwissenschaften sowie deren Lehr-Lernbedingungen erfasst, auf internationaler Ebene vergleicht und zeitliche Veränderungen in Deutschland seit 2007 berichtet. Die Umstellung des Erhebungsmodus in TIMSS 2019 ist durch eine international vergleichende Studie im Jahr 2017 untersucht worden, an der sich insgesamt 25 Länder beteiligt haben (Fishbein et al., 2018). Das Ziel dieser Vorstudie besteht in der bestmöglichen Übertragung der Testaufgaben aus TIMSS 2007, 2011 und 2015 von der papierbasierten in eine computerbasierte Form. Bei den Testaufgaben handelt es sich um sogenannte *Trenditems*, die zur Messung zeitlicher Veränderungen eingesetzt werden (zu den Herausforderungen bei unterschiedlichen Trendschätzungen vgl. Goldhammer et al., 2019; Robitzsch et al., 2020).

In TIMSS 2019 werden Testaufgaben und Schüler:innenantworten ebenso wie in früheren Zyklen nach der Item-Response-Theorie skaliert (Fishbein et al., 2018; vgl. auch Schwippert et al., 2020). Es wird also davon ausgegangen, dass durch die Testaufgaben eine zugrundeliegende Fähigkeit der Schüler:innen gemessen wird (d. h. Kompetenz in Mathematik und Naturwissenschaften). Fishbein und Kolleg:innen (2018) haben deshalb bereits im Rahmen der Moduseffektstudie untersucht, inwiefern bei der Erhebung der Schüler:innenkompetenzen durch CBT und PBT Unterschiede in den Testergebnissen oder Aufgabeneigenschaften bestehen. Wird durch beide Erhebungsmodi dieselbe zugrundeliegende Fähigkeit erfasst, so sollten CBT und PBT einerseits eine nahezu identische Rangreihung der Testergebnisse (Kompetenzwerte) ergeben. Andererseits sollten die Schwierigkeiten und Trennschärfen der Testaufgaben über die Erhebungsmodi hinweg weitestgehend identisch sein. Mit anderen Worten: Beobachtete Variabilität zwischen den durch PBT und CBT erfassten Schüler:innenantworten sollten zufällig zustande gekommen und nicht auf

systematische Unterschiede oder Verzerrungen zurückzuführen sein (Kane, 2013). Fishbein und Kolleg:innen (2018) zeigen zwar, dass die Schüler:innenkompetenzen im PBT und CBT in der Tat fast perfekt miteinander korrelieren, finden in Bezug auf die Aufgabenschwierigkeiten aber kleine Moduseffekte – ebenso wie andere Forscher:innen (Goldhammer et al., 2019; Robitzsch et al., 2020).

Diese Ergebnisse lassen sich so interpretieren, dass die Erfassung der mathematischen und naturwissenschaftlichen Schüler:innenkompetenzen in TIMSS 2019 nicht über Erhebungsmodi hinweg *generalisierbar* ist. Im Sinne eines aktuellen Verständnisses von Validität als „appropriateness, meaningfulness, and usefulness of specific inferences made from test scores" (AERA et al., 1999, S. 9) stellt die Generalisierbarkeit über verschiedene Testbedingungen hinweg jedoch eine wichtige Eigenschaft dar (Hartig, Frey & Jude, 2008; Kane, 2013), um verlässliche Aussagen über Schüler:innenkompetenzen treffen zu können. Wir gehen in diesem Beitrag also der Frage nach, inwieweit der Lösungserfolg bei der Aufgabenbearbeitung vom Erhebungsmodus abhängt.

Die Generalisierbarkeitstheorie (Cronbach et al., 1972) liefert dazu einen adäquaten methodologischen Rahmen. In einer G-Studie wird die beobachtete Variabilität in einer oder mehreren abhängigen Variablen im Hinblick auf verschiedene Varianzquellen zerlegt (z. B. Personen, Testaufgaben, Messzeitpunkte). Eine solche Varianzzerlegung erlaubt es, die Zuverlässigkeit eines Tests in Abhängigkeit von prinzipiell beliebig vielen interessierenden Testbedingungen zu beurteilen. Damit ist sie häufig informativer als eine Reliabilitätsschätzung nach der Klassischen Testtheorie (Shavelson & Webb, 1991). In aktuellen Arbeiten (z. B. Briggs & Wilson, 2007; Choi & Wilson, 2018) wird im Detail erläutert, wie sich Generalisierbarkeits- und Item-Response-Theorie gewinnbringend in der pädagogisch-psychologischen Forschungspraxis ergänzen können. In diesem Beitrag nutzen wir die methodologischen Überlegungen von Choi und Wilson (2018), um die Generalisierbarkeitstheorie auf kategoriale Daten (Schüler:innenantworten) anzuwenden.

Um potenzielle Moduseffekte bei der Bearbeitung der Testaufgaben aufzudecken, untersuchen wir außerdem, inwieweit deren Lösungshäufigkeit in PBT und CBT mit (äußeren) Aufgabenmerkmalen zusammenhängt. Zu den kognitiven Prozessen bei der Bearbeitung von Testaufgaben liegen reichhaltige psychologische und fachdidaktische Theorien vor, die empirische Lösungshäufigkeiten durch *schwierigkeitsgenerierende* Aufgabenmerkmale erklären (u. a. Antwortformat, Repräsentationsform; Gehrer, 2017; Prenzel, Häußler, Rost & Senkbeil, 2002). Solche Aufgabenmerkmale können den Lösungserfolg von Schüler:innen also systematisch und über interindividuelle Leistungsunterschiede hinaus beeinflussen (Hartig, 2007). Wenn mit beiden Erhebungsmodi dieselbe zugrundeliegende Fähigkeit erfasst wird, dann sollten schwierigkeitsgenerierende Merkmale der entsprechenden Testaufgaben ähnlich stark auf die Lösungshäufigkeit im PBT und CBT wirken. Mit anderen Worten: Der Erhebungsmodus sollte die Zusammenhänge zwischen schwierigkeitsgenerierenden Aufgabenmerkmalen und dem Lösungserfolg der Schüler:innen nicht moderieren (Goldhammer et al., 2019). In diesem Beitrag untersuchen wir

die folgenden Aufgabenmerkmale: die Inhaltsbereiche in Mathematik (Arithmetik, Messen/Geometrie, Daten) und Naturwissenschaften (Biologie, Physik/Chemie, Geografie), die kognitiven Anforderungsbereiche (Reproduzieren, Anwenden, Problemlösen), das Antwortformat (offen, geschlossen) sowie die Repräsentationsform (Text, Abbildung).

3. Methode

3.1 Untersuchungsdesign und Stichprobe

In Anlehnung an die internationalen Vorgaben (Fishbein et al., 2018) wurde eine Stichprobe von 895 Viertklässler:innen aus zwei Bundesländern gezogen (Bayern und Schleswig-Holstein). Davon haben 847 sowohl einen papierbasierten als auch einen computerbasierten Leistungstest in den Domänen Mathematik und Naturwissenschaften bearbeitet. Die Leistungstests wurden an zwei aufeinanderfolgenden Tagen durchgeführt: Die Schüler:innen bekamen entweder am ersten Testtag die papierbasierten und am zweiten Testtag die computerbasierten Leistungstests in beiden Domänen vorgelegt (PBT–CBT) oder durchliefen diese in umgekehrter Reihenfolge (CBT–PBT). Es nahmen insgesamt 23 Schulen mit jeweils zwei Parallelklassen an der Studie teil, wobei stets eine Klasse zufällig der Bedingung PBT–CBT ($n = 447$ Schüler:innen) und die andere der Bedingung CBT–PBT ($n = 400$ Schüler:innen) zugewiesen wurde. Das Design entspricht also einer Cross-over-Studie (u. a. Grizzle, 1965) mit je zwei Untersuchungsbedingungen (Reihenfolge der Testadministration) und Treatments (Erhebungsmodi). Cross-over-Studien haben gegenüber Versuchsplänen mit parallelen Treatmentgruppen (z. B. Schüler:innen bearbeiten entweder den papier- oder den computerbasierten Test) den Vorteil einer höheren Teststärke und den Nachteil möglicher Reihenfolgeeffekte. Beispielsweise könnten sich Lerneffekte einstellen oder Schüler:innen bei der Beantwortung des PBT weniger motiviert sein, wenn sie zuvor den CBT bearbeitet haben. Mit adäquaten Verfahren im Testdesign (vgl. Abschnitt 3.2) und bei der Datenanalyse können diese Effekte jedoch weitestgehend kontrolliert werden.

Ungefähr 52 % der Schüler:innen fühlen sich dem weiblichen Geschlecht zugehörig. Der Altersmedian beträgt 10 Jahre ($min = 9$, $max = 12$). Die Schüler:innen besitzen im Mittel (Median) „genug [Bücher], um ein Bücherregal zu füllen (26–100 Bücher)“ (fünfstufige Büchervariable, von *0–10 Bücher* bis *mehr als 200 Bücher*), was als Indikator für ihren sozioökonomischen Status dient (Wendt, Vennemann, Schwippert & Drossel, 2014). Die Schüler:innen wurden außerdem gebeten, ihre Fähigkeiten im Umgang mit digitalen Medien in zwei Bereichen einzuschätzen (jeweils vierstufige Antwortkategorien, Skalenbildung durch Summation). Einerseits wurde mit drei Items erfragt, inwieweit sie mit der Nutzung digitaler Medien vertraut sind (z. B. Touchscreens verwenden, $M = 7.13$, $SD = 1.90$, Cronbachs $\alpha = .73$). Andererseits wurden den Schüler:innen fünf computerbezogene Tätigkeiten vorgelegt. Für diese Tätigkeiten sollten die Schüler:innen angeben, inwieweit sie diese

ausführen können (z. B. Suchen von Informationen mit digitalen Medien, $M = 10.04$, $SD = 2.93$, Cronbachs $\alpha = .67$).

Durch die Reihenfolge der Testadministration ergeben sich in Bezug auf die Stichprobenmerkmale wie erwartet keine statistisch signifikanten Unterschiede (PBT–CBT vs. CBT–PBT, Geschlecht: 52 % vs. 53 % weiblich, adj. $F[1.0, 41.0] = 0.15$, $p = .699$, Altersmediane jeweils 10 Jahre, $min = 9$, $max = 12$, adj. $F[2.1, 85.1] = 0.16$, $p = .859$, Mediane der Büchervariablen jeweils *26–100 Bücher*, *min* = *0–10 Bücher*, *max* = *mehr als 200 Bücher*, adj. $F[5.5, 120.9] = 0.97$, $p = .441$, korrigierte Tests unter Berücksichtigung der geclusterten Datenstruktur nach Rao und Scott, 1987). Dies gilt ebenso für die selbsteingeschätzten Fähigkeiten der Schüler:innen (Vertrautheit: Cohens $d = -0.05$, $p = .506$; Tätigkeiten: Cohens $d = -0.06$, $p = .416$).

3.2 Leistungstests

Die Leistungstests bestehen aus mathematischen und naturwissenschaftlichen Testaufgaben, die in Aufgabenblöcken organisiert und mittels Rotationsschema zu Testheften zusammengefügt werden (Schwippert et al., 2020). In der Moduseffektstudie wurden insgesamt 214 Testaufgaben (Mathematik: 104, Naturwissenschaften: 110) und 16 Aufgabenblöcke verwendet (8 je Domäne) sowie acht verschiedene Testhefte erstellt. Dabei ist das Design der Testhefte für PBT und CBT identisch: Ein Testheft besteht aus je zwei Aufgabenblöcken in Mathematik und Naturwissenschaften. Jeder Aufgabenblock erscheint in mehreren Testheften an unterschiedlichen Positionen (mehr Informationen bei Fishbein et al., 2018). Durch ein solches Rotationsschema wird einerseits auf die Ausbalancierung von Positionseffekten (Mazzeo & von Davier, 2014) und andererseits auf die Vermeidung von Lerneffekten abgezielt, da die Schüler:innen verschiedene Testhefte im PBT und CBT und damit auch verschiedene Testaufgaben bearbeiten. Durch das Rotationsschema ist jede Testaufgabe in beiden Erhebungsmodi von knapp 200 Schüler:innen beantwortet worden. Insgesamt liegen damit je Domäne und Erhebungsmodus ungefähr 20 000 Schüler:innenantworten vor, da jedes Testheft etwa 25 Aufgaben enthält (847 × 25).

3.3 Aufgabenmerkmale

Verschiedene Studien konnten schwierigkeitsgenerierende Merkmale in Leistungstests ermitteln. Eine Untersuchung von Prenzel et al. (2002) mit den Daten der PISA-Naturwissenschaftsaufgaben identifiziert offene Antwortformate als in besonderer Weise schwierigkeitsgenerierend. Eine Studie im Rahmen des Nationalen Bildungspanels (NEPS) untersucht den Einfluss von Aufgaben- und Textmerkmalen auf die Itemschwierigkeit von Lesekompetenzaufgaben und bezieht dabei verschiedene Aufgaben- und Antwortformate im Allgemeinen (offen und geschlossen), die kognitiven Anforderungen der Items und verschiedene Textmerkmale mit ein (Gehrer, 2017). Jedem Item wurde also eine Kombination der genannten schwierigkeitsgene-

rierenden Aufgabenmerkmale für die Datenauswertung zugeordnet (Hartig, 2007). Im Folgenden werden diese überblicksartig beschrieben.

3.3.1 Inhaltsbereiche

Die Erfassung der Domänen Mathematik und Naturwissenschaften orientiert sich an den Curricula der an TIMSS teilnehmenden Staaten. Für die Domäne Mathematik korrespondiert der Leistungstest stark mit den 2004 in Deutschland festgelegten Bildungsstandards, in den Naturwissenschaften stellt der Perspektivrahmen der Gesellschaft für Didaktik des Sachunterrichts eine geeignete Orientierung dar (Steffensky, Scholz, Kasper & Köller, 2020). In Mathematik werden folgende Inhaltsbereiche erfasst: Arithmetik, Messen und Geometrie sowie Daten (Selter, Walter, Heinze, Brandt & Jentsch, 2020). Diese drei Inhaltsbereiche lassen sich weiterhin in verschiedene Teilgebiete untergliedern, so zählen zum Bereich Arithmetik etwa die Teilgebiete natürliche Zahlen, Terme, einfache Gleichungen und Beziehungen sowie Brüche und Dezimalzahlen (Selter et al., 2020). In den Naturwissenschaften werden folgende Inhalte abgedeckt: Biologie, Physik/Chemie und Geografie (Steffensky et al., 2020).

3.3.2 Kognitive Anforderungsbereiche (Reproduzieren, Anwenden, Problemlösen)

Die TIMSS-Items lassen sich unabhängig von der Domäne und den entsprechenden Inhaltsbereichen in Bezug auf zentrale Denkprozesse unterscheiden, die bei der Aufgabenbearbeitung aktiviert werden. Hierfür werden die folgenden drei kognitiven Anforderungsbereiche definiert: Reproduzieren, Anwenden (von Wissen, Fertigkeiten oder Grundvorstellungen) und Problemlösen (bei komplexen Aufgaben; Selter et al., 2020; Steffensky et al., 2020). Die Aufgaben verteilen sich in den Domänen Mathematik und Naturwissenschaften in ähnlicher Weise auf die kognitiven Anforderungsbereiche (Reproduzieren: 39 % vs. 42 %, Anwenden: 42 % vs. 35 %, Problemlösen: 19 % vs. 23 %).

3.3.3 Antwortformat (geschlossen, offen)

Die TIMSS-Leistungstests bestehen einerseits aus Multiple-Choice-Aufgaben (MC), bei denen eine oder mehrere Antworten ausgewählt werden müssen, und andererseits aus sogenannten Constructed-Response-Aufgaben (CR). Dabei handelt es sich um ein offenes Antwortformat, das eine kurze schriftliche Antwort von den Schüler:innen erfordert (Steffensky et al., 2020). Die Testaufgaben der Moduseffektstudie beinhalten sowohl in der Domäne Mathematik als auch in der Domäne Naturwissenschaften 45 % geschlossene (MC) und 55 % offene (CR) Aufgaben.

3.3.4 Repräsentationsform (Text, Abbildung)

Die TIMSS-Aufgaben wurden für die Datenauswertung gemäß ihrer Repräsentationsform kodiert. Hierbei wurde zwischen dem Vorhandensein von grafischen Elementen (z. B. Abbildungen oder Tabellen) und der Aufgabenformulierung in Textform ohne grafische Elemente unterschieden. Etwa die Hälfte der Mathematikaufgaben enthielt grafische Elemente (52 %), in den Naturwissenschaften war der Anteil etwas höher (61 %).

3.4 Statistische Analysen

Weil die abhängige Variable im vorliegenden Beitrag ein binäres Antwortformat aufweist (Schüler:innenantworten auf Testaufgaben sind 1 = richtig mit voller Punktzahl, andernfalls 0 = falsch), haben wir zur Bearbeitung unserer Forschungsfragen mehrere verallgemeinerte lineare gemischte Modelle berechnet. Diese erlauben einerseits die Schätzung zufälliger Effekte (Varianzkomponenten) für die G-Studien und andererseits die Spezifikation (fester) Haupt- und Interaktionseffekte, um die Zusammenhänge zwischen Erhebungsmodus, Aufgabenmerkmalen sowie dem Lösungserfolg der Schüler:innen zu bestimmen (vgl. auch De Boeck et al., 2011) und für die Untersuchungsbedingung (CBT–PBT vs. PBT–CBT) und den Testtag (1 vs. 2) zu kontrollieren.

Tabelle 1: Auflistung der untersuchten Varianzkomponenten mit Notation und Beschreibung

Varianzkomponente	Notation	Beschreibung
Schulen (s)	σ^2_s	Erwartete (mittlere) Unterschiede in den Testergebnissen von Viertklässler:innen an verschiedenen Schulen
Viertklässler:innen ($v{:}s$)	$\sigma^2_{v:s}$	Erwartete Unterschiede in den Testergebnissen verschiedener Viertklässler:innen (geclustert in Schulen)
Testaufgaben (t)	σ^2_t	Erwartete Unterschiede zwischen den Schwierigkeiten verschiedener Testaufgaben (Items)
Erhebungsmodi (m)	σ^2_m	Erwartete Unterschiede in den (mittleren) Testergebnissen von Viertklässler:innen in Abhängigkeit vom Erhebungsmodus
$t \times m$	σ^2_{tm}	Differenzielle Unterschiede zwischen den Schwierigkeiten der Testaufgaben in Abhängigkeit vom Erhebungsmodus, d. h. modusspezifische Unterschiede zwischen Aufgabenschwierigkeiten
$(v{:}s) \times t$	$\sigma^2_{(v:s)t}$	Differenzielle Unterschiede zwischen den Testergebnissen der Viertklässler:innen in Abhängigkeit von der Testaufgabe
$(v{:}s) \times m$	$\sigma^2_{(v:s)m}$	Differenzielle Unterschiede zwischen den Testergebnissen der Viertklässler:innen in Abhängigkeit vom Erhebungsmodus
Residuum (e)	σ^2_e	Unaufgeklärter Varianzanteil, der auf die Interaktionseffekte zwischen *allen* berücksichtigten Testbedingungen (Interaktion höchster Ordnung) und/oder nicht beobachtete Variablen zurückgeht

In Tabelle 1 sind die Varianzkomponenten aufgeführt, die in den G-Studien berücksichtigt werden. Die Variabilität in den Antworten der Schüler:innen zerfällt also in Schuleffekte (*s*), Unterschiede zwischen Schüler:innen (*v:s*), Unterschiede zwischen den Schwierigkeiten der Testaufgaben (*t*) und Erhebungsmodi (*m*) sowie Interaktionen zwischen diesen Testbedingungen. Unterschiede, die auf den Erhebungsmodus zurückzuführen sind, haben wir in den Modellen zunächst als zufällige Effekte spezifiziert, um die Generalisierbarkeit über Testbedingungen hinweg zu untersuchen (Cronbach et al., 1972; Kane, 2013). Da bisherige Befunde allerdings eher für systematische Unterschiede zwischen den Erhebungsmodi (feste Effekte) sprechen, haben wir zusätzlich G-Studien durchgeführt, die ohne Varianzkomponenten für Moduseffekte auskommen. Feste Effekte werden in G-Studien normalerweise nicht explizit berücksichtigt, da sie die Reliabilität einer Messung im Sinne der G-Theorie nicht beeinflussen (Cronbach et al., 1972). Um Hinweise auf eine erfolgreiche Randomisierung zu finden, führen wir getrennte G-Studien für PBT–CBT und CBT–PBT durch.

Auf die in der Literatur vorgeschlagene detaillierte Reliabilitätsschätzung in G-Studien (z. B. Shavelson & Webb, 1991) verzichten wir in diesem Beitrag mit folgendem Grund: Die Bestimmung von Messfehler und Reliabilität ist in einer G-Studie abhängig von den Testbedingungen, über die eine Generalisierbarkeit der Testwerte angenommen wird *und* der Anzahl der Beobachtungen in der abhängigen Variablen. Mit anderen Worten: Aus einer einzigen G-Studie lassen sich zahlreiche Messfehler und Reliabilitäten bestimmen. Dies hängt damit zusammen, dass die Generalisierbarkeit der Testwerte immer in Bezug auf ein bestimmtes Ziel untersucht wird (z. B. absolute oder relative Entscheidungen; Cronbach et al., 1972). In der vorliegenden Untersuchung ist dieses Potenzial von G-Studien kaum bedeutsam, weil keine Unsicherheit in Bezug auf die Zielbestimmung besteht.[1]

Wir steigern die Modellkomplexität sukzessive, indem wir zunächst Modelle schätzen, die ausschließlich Varianzkomponenten gemäß Tabelle 1 enthalten. Wir erweitern diese Modelle dann durch feste Effekte für (a) den Erhebungsmodus und die Kontrollvariablen (Untersuchungsbedingung, Testtag), (b) Aufgabenmerkmale und schließlich (c) Interaktionen zwischen Aufgabenmerkmalen und dem Erhebungsmodus. Für die Modellvergleiche nutzen wir die Informationskriterien AIC und BIC, bei denen ein niedrigerer Wert eine bessere Modellanpassung bedeutet. Regressionskoeffizienten geben wir als Odds Ratios (*OR*, Chancenverhältnisse) und in der gängigen Metrik als logarithmierte Chancenverhältnisse an (unstandardisierter Regressionskoeffizient). Für statistische Tests verwenden wir ein Signifikanzniveau von $p = .05$. Alle nachfolgenden Berechnungen wurden in R (R Core Team,

1 Außerdem liegt eine sehr große Anzahl an Beobachtungen für die abhängige Variable vor. Eine Reliabilitätsschätzung würde daher unabhängig vom Erhebungsmodus stets ein (fast) perfektes Ergebnis liefern. In der Tat erhalten wir unter Verwendung der Formeln von Choi und Wilson (2018) Werte von über .90 für beide Modi, Untersuchungsbedingungen und Domänen.

2022) mit dem Paket lme4 (Bates, Mächler, Bolker & Walker, 2015) durchgeführt (vgl. Auswertungssyntax unter https://osf.io/k25y7).[2]

4. Ergebnisse

4.1 Generalisierbarkeitsstudien

Die Ergebnisse der G-Studien für die Domänen Mathematik und Naturwissenschaften sind in den Tabellen 2 (zufällige Moduseffekte) und 3 (feste Moduseffekte) dargestellt. Wir haben separate G-Studien für die beiden Untersuchungsbedingungen PBT–CBT und CBT–PBT durchgeführt. Die Ergebnisse fallen insgesamt ähnlich aus und zwar unabhängig von der Domäne, der Spezifikation der Moduseffekte oder der Untersuchungsbedingung.

Der Anteil der Varianz, der auf Unterschiede zwischen den teilnehmenden Schulen zurückgeführt werden kann, ist zumeist gering (max. 4.5 %, meist < 4 %) und variiert in den Naturwissenschaften etwas stärker zwischen den Untersuchungsbedingungen PBT–CBT und CBT–PBT als in Mathematik. Auf Unterschiede zwischen Schüler:innen entfällt ein Varianzanteil von etwa 10 % (Naturwissenschaften) bis 15 % (Mathematik). Im Sinne Cohens (1992) kann dies zwar als moderate Effektstärke verstanden werden, da die Leistungstests jedoch auf interindividuelle Unterschiede zwischen Schüler:innen abzielen, sind die Varianzanteile in Bezug auf die Konstruktvalidität tendenziell kritisch zu beurteilen.

Der Varianzanteil, der auf (Schwierigkeits-)Unterschiede zwischen Testaufgaben entfällt, liegt für alle Modelle und Domänen bei etwa einem Viertel der Gesamtvarianz. In keinem Fall findet sich ein Interaktionseffekt zwischen Schüler:innen und Items, der dafür spräche, dass der Lösungserfolg bei einer Testaufgabe zwischen den Schüler:innen variieren würde (Varianzanteil < 1 %). Das kann so interpretiert werden, dass die *Trennschärfe* der eingesetzten Items ähnlich ausfällt. Der größte Varianzanteil bleibt in allen G-Studien unaufgeklärt (etwa 60 %). Dies deutet darauf hin, dass wichtige Testbedingungen bei der Analyse nicht berücksichtigt werden konnten.

Wir bemerken abschließend: (a) Mit Ausnahme des Schuleffekts fallen die Befunde aus den G-Studien für beide Untersuchungsbedingungen (Reihenfolge der Testadministration) fast identisch aus, wie dies nach einer randomisierten Zuweisung auch zu erwarten wäre; (b) zufällige Moduseffekte vermögen in allen Modellen nur einen marginalen Varianzanteil zu erklären (insgesamt weniger als 2.5 % der Gesamtvarianz, vgl. Tabelle 2). In den nachfolgenden gemischten Modellen gehen wir deshalb von festen Effekten für die Untersuchungsbedingungen und Erhebungsmodi aus.

---

2 Zur Generierung der Datensätze und um die in Abschnitt 3.1 berichteten Tests auf Unterschiede zwischen den Untersuchungsbedingungen durchzuführen, wurde IBM SPSS 28 verwendet.

Tabelle 2: Ergebnisse der G-Studien mit zufälligen Moduseffekten in Abhängigkeit von Domänen und Untersuchungsbedingungen

	Mathematik						Naturwissenschaften					
	PBT–CBT		CBT–PBT		Total		PBT–CBT		CBT–PBT		Total	
Varianzkomponente	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%
σ_s^2	0.11	2.0	0.24	4.2	0.16	2.8	0.02	0.4	0.25	4.5	0.08	1.5
$\sigma_{v:s}^2$	0.76	13.7	0.82	14.2	0.80	14.2	0.51	9.6	0.56	10.0	0.54	9.8
σ_t^2	1.26	22.7	1.30	22.6	1.28	22.7	1.39	26.1	1.38	24.6	1.50	27.2
σ_m^2	0.03	0.5	0.02	0.4	0.02	0.4	0.02	0.4	0.02	0.4	0.02	0.4
σ_{tm}^2	0.04	0.7	0.04	0.7	0.04	0.7	0.04	0.8	0.04	0.7	0.03	0.5
$\sigma_{(v:s)t}^2$	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
$\sigma_{(v:s)m}^2$	0.05	0.9	0.05	0.9	0.05	0.9	0.05	0.9	0.06	1.1	0.05	0.9
σ_e^2	3.29	59.4	3.29	57.1	3.29	58.3	3.29	61.8	3.29	58.8	3.29	59.7
Gesamtvarianz	5.54	100.0	5.76	100.0	5.64	100.0	5.32	100.0	5.60	100.0	5.51	100.0

Anmerkungen. 23 Schulen, 847 Schüler:innen, 104 Testaufgaben in Mathematik, 110 Testaufgaben in Naturwissenschaften und zwei Erhebungsmodi (PBT vs. CBT).

Tabelle 3: Ergebnisse der G-Studien mit festen Moduseffekten in Abhängigkeit von Domänen und Untersuchungsbedingungen

	Mathematik						Naturwissenschaften					
	PBT–CBT		CBT–PBT		Total		PBT–CBT		CBT–PBT		Total	
Varianzkomponente	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%	Schätzer	%
σ_s^2	0.11	2.0	0.23	4.1	0.15	2.7	0.02	0.4	0.15	2.8	0.08	1.5
$\sigma_{v:s}^2$	0.77	14.3	0.83	14.7	0.80	14.6	0.53	10.2	0.57	10.7	0.55	10.2
σ_t^2	1.23	22.8	1.29	22.9	1.26	22.9	1.38	26.4	1.34	25.1	1.49	27.5
$\sigma_{(v:s)t}^2$	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0	0.00	0.0
σ_e^2	3.29	60.9	3.29	58.3	3.29	59.8	3.29	63.0	3.29	61.5	3.29	60.8
Gesamtvarianz	5.40	100.0	5.64	100.0	5.50	100.0	5.22	100.0	5.35	100.0	5.41	100.0

Anmerkungen. 23 Schulen, 847 Schüler:innen, sowie 104 Testaufgaben in Mathematik und 110 Testaufgaben in Naturwissenschaften.

4.2 Zusammenhänge zwischen Aufgabenmerkmalen und Schüler:innenantworten

Gemischte Modelle mit Varianzkomponenten für Schulen, Schüler:innen und Testaufgaben sowie festen Effekten für den Modus, die Untersuchungsbedingung und den Testtag deuten an, dass der computerbasierte Erhebungsmodus negativ mit dem Lösungserfolg in den Testaufgaben in Mathematik und in den Naturwissenschaften zusammenhängt (Mathematik: $B = -0.24$, $SE = 0.02$, $OR = 0.79$, $p < .001$; Naturwissenschaften: $B = -0.24$, $SE = 0.02$, $OR = 0.78$, $p < .001$). Die Untersuchungsbedingung hat erwartungsgemäß keinen Einfluss auf den Lösungserfolg der Schüler:innen (Mathematik: $B = 0.01$, $SE = 0.07$, $OR = 1.00$, $p = .952$; Naturwissen-

Tabelle 4: Effekte der Erhebungsmodi und Aufgabenmerkmale auf den Lösungserfolg in Mathematik

	Modell M1			Modell M2		
	Schätzer	*SE*	*p*	Schätzer	*SE*	*p*
(Konstante)	1.44	0.26	< .001	1.33	0.27	< .001
Bedingung: CBT–PBT	–0.01	0.07	.892	–0.01	0.07	.896
Modus: CBT[a]	–0.22	0.02	< .001	–0.14	0.05	.006
Inhaltsbereiche:[b]						
Messen/Geometrie	0.73	0.41	.071	0.94	0.43	.029
Daten	0.21	0.30	.488	0.49	0.30	.116
Anforderungsbereiche:[c]						
Anwenden	–1.05	0.24	< .001	–1.11	0.25	< .001
Problemlösen	–0.62	0.30	.040	–0.37	0.31	.241
Antwortformat: Offen[d]	–0.15	0.22	.488	–0.07	0.23	.752
Repräsentation: Abb./Tab.[e]	–0.43	0.27	.119	–0.53	0.29	.071
Interaktionseffekte						
Modus × Messen/Geometrie				–0.14	0.09	.149
Modus × Daten				–0.19	0.07	.006
Modus × Anwenden				–0.04	0.06	.464
Modus × Problemlösen				–0.16	0.07	.021
Modus × Antwortformat				–0.05	0.05	.326
Modus × Repräsentationsform				0.07	0.07	.304
Varianzkomponenten						
Schulen			0.16			0.16
Schüler:innen			0.75			0.75
Testaufgaben			1.05			1.05

Anmerkungen. Es wurde für den Testtag (fester Effekt) und differenzielle Unterschiede zwischen Schüler:innen in Abhängigkeit von der Testaufgabe kontrolliert (Varianzkomponente $[v:s] \times t$).
[a] Referenzkategorie ist der papierbasierte Test (PBT). [b] Referenzkategorien sind Arithmetik in Mathematik und Biologie in Naturwissenschaften. [c] Referenzkategorie ist der Anforderungsbereich Reproduzieren. [d] Referenzkategorie ist das Antwortformat geschlossen. [e] Referenzkategorie ist die Repräsentationsform Text.

Tabelle 5: Effekte der Erhebungsmodi und Aufgabenmerkmale auf den Lösungserfolg in Naturwissenschaften

	Modell N1			Modell N2		
	Schätzer	*SE*	*p*	Schätzer	*SE*	*p*
(Konstante)	0.82	0.26	< .001	0.69	0.27	.013
Bedingung: CBT–PBT	0.00	0.06	.999	–0.00	0.06	.978
Modus: CBT[a]	–0.24	0.02	< .001	–0.15	0.06	.006
Inhaltsbereiche:[b]						
Physik/Chemie	0.31	0.24	.191	0.25	0.26	.331
Geografie	0.18	0.25	.471	0.24	0.27	.372
Anforderungsbereiche:[c]						
Anwenden	–0.13	0.24	.588	0.02	0.26	.939
Problemlösen	–0.53	0.27	.048	–0.61	0.29	.035
Antwortformat: Offen[d]	0.12	0.20	.551	0.20	0.21	.341
Repräsentation: Abb./Tab.[e]	0.14	0.22	.511	0.26	0.23	.258
Interaktionseffekte						
Modus × Physik/Chemie				0.04	0.06	.445
Modus × Geografie				–0.04	0.06	.549
Modus × Anwenden				–0.10	0.06	.092
Modus × Problemlösen				0.05	0.06	.408
Modus × Antwortformat				–0.05	0.05	.277
Modus × Repräsentationsform				–0.08	0.05	.137
Varianzkomponenten						
Schulen			0.08			0.08
Schüler:innen			0.96			0.96
Testaufgaben			0.56			0.56

Anmerkungen. Es wurde für den Testtag (fester Effekt) und differenzielle Unterschiede zwischen Schüler:innen in Abhängigkeit von der Testaufgabe kontrolliert (Varianzkomponente [*v*:*s*] × *t*).
[a] Referenzkategorie ist der papierbasierte Test (PBT). [b] Referenzkategorien sind Arithmetik in Mathematik und Biologie in Naturwissenschaften. [c] Referenzkategorie ist der Anforderungsbereich Reproduzieren. [d] Referenzkategorie ist das Antwortformat geschlossen. [e]Referenzkategorie ist die Repräsentationsform Text.

schaften: $B = 0.00$, $SE = 0.06$, $OR = 1.00$, $p = .928$). Wir haben schließlich auch geprüft, inwieweit sich Lern- oder Ermüdungseffekte eingestellt haben, der Lösungserfolg in den Testaufgaben also vom Testtag abhängt. Dies ist allerdings nicht der Fall (Mathematik: $B = -0.05$, $SE = 0.02$, $OR = 0.95$, $p = .051$; Naturwissenschaften: $B = -0.02$, $SE = 0.02$, $OR = 0.98$, $p = .338$).

In den Tabellen 4 und 5 sind die Ergebnisse gemischter Modelle mit festen Haupt- (M1, N1) und Interaktionseffekten (M2, N2) für die Aufgabenmerkmale dargestellt. Die Modelle *ohne* Interaktionseffekte zeigen eine deutlich bessere Anpassung an die Daten als die zuvor berichteten Basismodelle (M0, N0) und je nach Kriterium eine ähnliche oder etwas bessere als die komplexeren Modelle M2 und N2 (Mathematik: M0 vs. M1 vs. M2, AIC = 44475 vs. 41331 vs. 41324, BIC = 44544 vs.

41451 vs. 41495; Naturwissenschaften: N0 vs. N1 vs. N2, AIC = 45249 vs. 44272 vs. 44272, BIC = 45318 vs. 44392 vs. 44443).

Für beide Domänen findet sich auch in M1 und N1 ein kleiner Moduseffekt (Mathematik: $OR = 0.80$, $p < .001$; Naturwissenschaften: $OR = 0.79$, $p < .001$) derart, dass der Lösungserfolg der Schüler:innen im CBT niedriger ist als im PBT. In den komplexeren Modellen M2 und N2 ist der Effekt etwas kleiner, aber weiterhin negativ (Mathematik: $OR = 0.87$, $p = .006$; Naturwissenschaften: $OR = 0.86$, $p = .006$). Für Mathematik ergibt sich in M1 außerdem ein negativer Effekt der kognitiven Anforderungsbereiche Anwenden ($OR = 0.35$, $p < .001$) und Problemlösen ($OR = 0.54$, $p = .040$) gegenüber dem Referenzbereich Reproduzieren. Schüler:innen lösen Testaufgaben also (erwartungsgemäß) mit deutlich höherer Wahrscheinlichkeit, wenn diese aus dem Anforderungsbereich Reproduzieren stammen. Modell M1 zeigt für die Domäne Mathematik keine weiteren Haupteffekte. In Modell M2 findet sich zusätzlich ein positiver Haupteffekt für den Inhaltsbereich Geometrie ($OR = 2.56$, $p = .029$), der andeutet, dass Schüler:innen entsprechende Aufgaben mit einer mehr als doppelt so hohen Wahrscheinlichkeit lösen wie solche im Bereich Arithmetik. Ferner wird der negative Effekt für den Anforderungsbereich Problemlösen durch eine Interaktion mit dem Erhebungsmodus spezifiziert ($OR = 0.85$, $p = .021$). Das gilt in ähnlicher Weise für Aufgaben aus dem Inhaltsbereich Daten ($OR = 0.83$, $p = .006$). Beide deuten an, dass der CBT in Mathematik Schüler:innen zusätzliche Schwierigkeiten in den genannten Bereichen bereitet. Weitere Interaktionseffekte finden sich nicht. Für die Naturwissenschaften findet sich nur ein statistisch signifikanter Zusammenhang zwischen dem kognitiven Anforderungsbereich Problemlösen und dem Lösungserfolg der Schüler:innen bei der Aufgabenbearbeitung (N1: $OR = 0.59$, $p = .048$; N2: $OR = 0.54$, $p = .048$). Demnach fällt es Schüler:innen gegenüber dem Referenzbereich Reproduzieren (wiederum erwartungsgemäß) schwerer, entsprechende Aufgaben zu lösen.[3]

5. Diskussion

Die Umstellung auf eine computerbasierte Erfassung von Schüler:innenkompetenzen in internationalen Vergleichsstudien ist nicht nur eine Reaktion auf die stetig zunehmende Digitalisierung in unserer Gesellschaft, sondern hat auch Vorteile in der Datenerhebung und -verarbeitung (z. B. Frey & Hartig, 2013). Die TIMSS 2019 vorgelagerte Moduseffektstudie hatte deshalb das Ziel, die Testaufgaben, die zur Messung zeitlicher Veränderungen in den Schüler:innenkompetenzen eingesetzt werden, bestmöglich von PBT zu CBT zu übertragen und diesen Prozess zu evaluieren. Eine solche Übertragung der Leistungstests erhebt den Anspruch der Generali-

3 Um die Robustheit unserer Ergebnisse zu überprüfen, haben wir abschließend noch separate gemischte Modelle für jedes Aufgabenmerkmal berechnet. Dabei haben sich lediglich marginale Abweichungen in den Chancenverhältnissen gegenüber den in den Tabellen 4 und 5 berichteten ergeben. Insbesondere haben sich keine weiteren statistisch signifikanten Moderatoreffekte für den Erhebungsmodus gezeigt.

sierbarkeit über Erhebungsmodi hinweg. Bisherige Studien im Rahmen von TIMSS 2019 (Fishbein et al., 2018; Robitzsch et al., 2020) und PISA 2018 (Goldhammer et al., 2019) liefern empirische Evidenz, die zumindest in Bezug auf die Rangreihung der Schüler:innenkompetenzen *für* diese Generalisierbarkeit sprechen, da eine Verzerrung der Rangreihung durch die Umstellung des Erhebungsmodus anscheinend nicht stattfindet. Die Studien berichten jedoch – ebenso wie dieser Beitrag – kleine Moduseffekte auf die Aufgabenschwierigkeiten: Testaufgaben haben zumeist höhere Lösungswahrscheinlichkeiten, wenn sie den Schüler:innen im papierbasierten Format präsentiert werden, und zwar sowohl in Mathematik als auch in den Naturwissenschaften.

In diesem Beitrag sind im Rahmen einer randomisierten Cross-over-Studie zwei weiterführende Forschungsfragen bearbeitet worden, die die Generalisierbarkeit der Erfassung von mathematischen und naturwissenschaftlichen Schüler:innenkompetenzen durch CBT und PBT thematisieren. Wir haben dafür zunächst eine Zerlegung der Varianz in den Schüler:innenantworten durchgeführt, um festzustellen, ob diese durch modusabhängige Unterschiede in der Aufgabenbearbeitung erklärt werden können. Des Weiteren haben wir untersucht, inwieweit der Erhebungsmodus die Zusammenhänge zwischen Aufgabenmerkmalen und dem Lösungserfolg der Schüler:innen moderiert. Ein solcher Befund würde *gegen* eine generalisierbare Erfassung der Schüler:innenkompetenzen sprechen, da dieser als Hinweis für modusabhängige Unterschiede bei der Aufgabenbearbeitung gedeutet werden müsste (Goldhammer et al., 2019). Unsere Ergebnisse zeigen allerdings, dass der Erhebungsmodus sowohl in Bezug auf Haupt- als auch auf Interaktionseffekte kaum Varianz in den Schüler:innenantworten erklärt. In Bezug auf die Effekte schwierigkeitsgenerierender Merkmale auf die Bearbeitung von Mathematikaufgaben finden sich zwar statistisch signifikante Unterschiede für die Bereiche Daten und Problemlösen, diese sind allerdings klein. Für die Naturwissenschaften lassen sich diesbezüglich keine Interaktionseffekte nachweisen.

Die praktische Bedeutsamkeit von Moduseffekten für Schulleistungsuntersuchungen wie PISA oder TIMSS ist damit im Lichte der vorliegenden Befunde nur von marginaler Bedeutung: In komplexen, aufwändig kontrollierten Studien können solche Verzerrungen berücksichtigt und relativ problemlos durch die Addition einer geeigneten Konstanten oder durch Ausschluss nicht invarianter Testaufgaben korrigiert werden (Fishbein et al., 2018; Robitzsch et al., 2020; Schwippert et al., 2020). Die Umstellung des Erhebungsmodus in TIMSS kann insofern als gelungen bezeichnet werden, als das Ziel der internationalen Studienleitung die äquivalente Erfassung der Schüler:innenkompetenzen durch CBT und PBT in den beiden Domänen Mathematik und Naturwissenschaften war. Die vorgelegten Befunde unterstützen diese Aussage weitestgehend, sodass zukünftig die Vorteile computerbasierter Leistungstests stärker genutzt werden könnten (u. a. Cotter et al., 2020).

Das Ergebnis, dass insbesondere für den mathematischen Leistungstest Moduseffekte zugunsten von PBT gefunden wurden, dürfte für die unterrichtliche Praxis in Deutschland jedoch durchaus bedeutsam sein. In Bezug auf das The-

ma dieses Beitrags lässt sich vermuten, dass Schüler:innen im computerbasierten Leistungstest zusätzlich zu ihrem Wissen in der jeweiligen Domäne von Fähigkeiten profitieren könnten, die durch den routinierten Umgang mit digitalen Medien im Unterricht trainiert werden. Diese Routine ist bei den Viertklässler:innen in der vorliegenden Studie nur teilweise ausgeprägt. Die unterrichtliche oder private Nutzung digitaler Medien ist allerdings nicht per se förderlich für die Wissensentwicklung der Schüler:innen. Stattdessen sollten diese gezielt digitale Kompetenzen erwerben, um Lernprozesse in verschiedenen, teils herausfordernden Kontexten erfolgreich bewältigen zu können (z. B. Homeschooling; Eickelmann, Bos, Gerick & Labusch, 2019; Schaumburg, Gerick, Eickelmann & Labusch, 2019).

Abschließend soll noch auf einige Limitationen dieser Studie hingewiesen werden. Erstens berichten wir Daten aus lediglich zwei Ländern der Bundesrepublik Deutschland. Die (externe) Generalisierbarkeit der vorliegenden Ergebnisse kann deshalb zum Anlass für weitere Studien genommen werden, wobei mit der Haupterhebung von TIMSS 2019 entsprechende Daten und Ergebnisse (Robitzsch et al., 2020) vorliegen. Zweitens blieben die Zusammenhänge zwischen den Schüler:innenantworten und den im Allgemeinen als schwierigkeitsgenerierend angenommenen Merkmalen der Testaufgaben zumeist aus (vgl. auch Walzebug, 2015). Dies könnte darauf hindeuten, dass es andere, möglicherweise besser geeignete Aufgabenmerkmale gibt, die bei der Bearbeitung der Leistungstests bedeutsam sind (z. B. kognitives Aktivierungspotenzial; Maier, Kleinknecht, Metz & Bohl, 2010).

Da entsprechende Zusammenhänge aus früheren Studien gut dokumentiert sind (Gehrer, 2017; Hartig, 2007; Stahns, Walzebug & Kasper, 2016), vermuten wir jedoch eher, dass die hier untersuchten Merkmale nicht sensitiv genug für die vorliegende Aufgabenstichprobe waren. Beispielsweise könnte man annehmen, dass sich die Testaufgaben in der Repräsentationsform oder im Antwortformat (mit gutem Grund) nicht stark genug voneinander unterscheiden, um Effekte auf die Aufgabenschwierigkeit zu finden. Vielleicht wäre es deshalb gewinnbringend, die Testadministration im CBT verstärkt in den Blick zu nehmen, um die zugrundeliegenden kognitiven Prozesse der Schüler:innen besser beschreiben zu können. Kröhne und Martens (2011) diskutieren einige Eigenschaften der Testadministration, die zu einer erhöhten Schwierigkeit im CBT führen können (z. B. Nutzung von Maus und Tastatur, Anzahl der Items pro angezeigter Seite). Dabei spielt die Bearbeitungszeit im CBT eine wichtige moderierende Rolle (Mead & Drasgow, 1993), die in dieser Studie nicht in den Blick genommen wurde.

Literatur

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bos, W., Eickelmann, B., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K. … Wendt, H. (Hrsg.). (2014). *ICILS 2013. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern in der 8. Jahrgangsstufe im internationalen Vergleich*. Waxmann.

Briggs, D. C. & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*(2), 131–155. https://doi.org/10.1111/j.1745-3984.2007.00031.x

Choi, J. & Wilson, M. R. (2018). Modeling rater effects using a combination of Generalizability Theory and IRT. *Psychological Test and Assessment Modeling, 60*(1), 53–80.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cotter, K. E., Centurino, V. A. S. & Mullis, I. V. S. (2020). Developing the TIMSS 2019 mathematics and science achievement instruments. In M. O. Martin, M. von Davier & I. V. S. Mullis (Hrsg.), *Methods and procedures: TIMSS 2019 technical report* (S. 1.1–1.36). TIMSS & PIRLS International Study Center, Boston College.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F. & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28. https://doi.org/10.18637/jss.v039.i12

Eickelmann, B., Bos, W., Gerick, J. & Labusch, A. (2019). Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern der 8. Jahrgangsstufe in Deutschland im zweiten internationalen Vergleich. In B. Eickelmann, W. Bos, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert … J. Vahrenhold (Hrsg.), *ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking* (S. 113–136). Waxmann.

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 implication for measuring. *Large-scale Assessments in Education, 6*(1), Article 11. https://doi.org/10.1186/s40536-018-0064-z

Frey, A. & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von papier- und bleistift-basierten Verfahren eingesetzt werden? In D. Leutner, E. Klieme, J. Fleischer & H. Kuper (Hrsg.), *Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen* (Zeitschrift für Erziehungswissenschaft, Sonderheft 18, S. 53–57). https://doi.org/10.1007/s11618-013-0385-1

Gehrer, K. (2017). *Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben* (NEPS Working Paper No. 67). Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Goldhammer, F., Harrison, S., Bürger, S., Kroehne, U., Lüdtke, O., Robitzsch, A. … 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: Grundbildung im internationalen Vergleich* (S. 163–186). Waxmann.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics, 21*(2), 467–480. https://doi.org/10.2307/2528104

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Beltz.

Hartig, J., Frey, A. & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 135–163). Springer. https://doi.org/10.1007/978-3-540-71635-8_7

Kane, M.T. (2013) Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the national education panel study: The challenge of mode effects. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Hrsg.), *Education as a lifelong process* (Zeitschrift für Erziehungswissenschaft, Sonderheft 14, S. 169–186). https://doi.org/10.1007/s11618-011-0185-4

Maier, U., Kleinknecht, M., Metz, K. & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung, 28*(1), 84–96. https://doi.org/10.36950/bzl.28.1.2010.9798

Mazzeo, J. & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis* (S. 229–258). CRC Press.

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*(2), 120–135.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rao, J.N.K. & Scott, A.J. (1987). On simple adjustments to chi-square tests with survey data. *Annals of Statistics, 15*(1), 385–397. https://doi.org/10.1214/aos/1176350273

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.). (2016). *PISA 2015. Eine Studie zwischen Kontinuität und Innovation*. Waxmann.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F. & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien: Eine Skalierung der deutschen PISA-Daten. *Diagnostica, 63*(2), 148–165. https://doi.org/10.1026/0012-1924/a000177

Robitzsch, A., Lüdtke, O., Schwippert, K., Goldhammer, F., Kroehne, U. & Köller, O. (2020). Leistungsveränderungen in TIMSS zwischen 2015 und 2019: Die Rolle des Testmediums und des methodischen Vorgehens bei der Trendschätzung. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 169–186). Waxmann.

Schaumburg, H., Gerick, J., Eickelmann, B. & Labusch, A. (2019). Nutzung digitaler Medien aus der Perspektive der Schülerinnen und Schüler im internationalen Vergleich. In B. Eickelmann, W. Bos, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert ... & J. Vahrenhold (Hrsg.), *ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking* (S. 241–270). Waxmann.

Schwippert, K., Scholz, L.A., Beese, C., Kasper, D., Schulz-Heidorf, K. & Girelli, A.-L. (2020). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2019). In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 25–56). Waxmann. https://doi.org/10.31244/9783830993193

Selter, C., Walter, D., Heinze, A., Brandt, J. & Jentsch, A. (2020). Mathematische Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im Vergleich* (S. 57–113). Waxmann.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Stahns, R., Walzebug, A. & Kasper, D. (2016). (Bildungs-)sprachliche Anforderungen und Aufgabenschwierigkeit in Sachtext-Leseitems der Internationalen Grundschul-Lese-Untersuchung (IGLU 2011). In R. Strietholt, W. Bos, H.G. Holtappels & N. McElvany (Hrsg.), *Jahrbuch der Schulentwicklung* (Bd. 19, S. 57–83). Beltz Juventa.

Steffensky, M., Scholz, L.A., Kasper, D. & Köller, O. (2020). Naturwissenschaftliche Kompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im Vergleich* (S. 115–168). Waxmann.

Walzebug, A. (2015). *Sprachlich bedingte Ungleichheit. Theoretische und empirische Betrachtungen am Beispiel mathematischer Testaufgaben und ihrer Bearbeitung*. Waxmann.

Wang, S., Jiao, H., Young, M.J., Brooks, T. & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238. https://doi.org/10.1177/0013164406288166

Wendt, H., Vennemann, M., Schwippert, K. & Drossel, K. (2014). Soziale Herkunft und computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im internationalen Vergleich. In W. Bos, B. Eickelmann, J. Gerick, F. Goldhammer, H. Schaumburg, K. Schwippert … H. Wendt (Hrsg.), *ICILS 2013. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern in der 8. Jahrgangsstufe im internationalen Vergleich* (S. 265–296). Waxmann.

Zinn, S., Landrock, U. & Gnambs, T. (2021). Web-based and mixed-mode cognitive large-scale assessments in higher education: An evaluation of selection bias, measurement bias, and prediction bias. *Behavior Research Methods, 53*(3), 1202–1217. https://doi.org/10.3758/s13428-020-01480-7