Doris Lewalter • Jennifer Diedrich
Frank Goldhammer • Olaf Köller
Kristina Reiss (Hrsg.)

PISA 2022

Analyse der Bildungsergebnisse in Deutschland

Online-Kapitel 12

WAXMANN

Inhalt

12 Dokumentation der Methoden zur Datenerhebung und Auswertung für PISA 2022 in Deutschland

Zentrum für internationale Bildungsvergleichsstudien (ZIB)
& IEA-Hamburg

Die Durchführung der PISA-Studien in Deutschland ist ein methodisch und organisatorisch anspruchsvolles Unterfangen, das eine enge Zusammenarbeit von unterschiedlichen Personen in Schulen, IEA-Hamburg, internationalen Partnern und Vertragsnehmern der OECD sowie dem Projektzentrum an der TU in München erfordert. Die erfolgreiche Durchführung der notwendigen Vorbereitungen, der Datenerhebung im Feld und die sich anschließenden Weiterverarbeitungsprozesse bedürfen einer sorgfältigen Planung und Koordination auf verschiedenen Ebenen. In Deutschland sind Schulen aus allen Bundesländern beteiligt, und es ist wichtig, dass die PISA-Testung in allen Schulen nach den gleichen Standards und Richtlinien abläuft. Besonders herausfordernd war die Durchführung von PISA 2022 in Deutschland wegen der Coronavirus-Pandemie (SARS-CoV-2). Die Belastungen und Einschränkungen, die mit dem pandemischen Geschehen einhergingen, führten zu zusätzlichen Schwierigkeiten bei der Organisation und Durchführung der Tests. Schulschließungen, Wechselunterricht, Hygienevorschriften und nicht zuletzt krankheitsbedingte Ausfälle erschwerten eine fixe Planung und Umsetzung der Tests, und es waren zusätzliche Maßnahmen erforderlich, um die Sicherheit der beteiligten Personen zu gewährleisten. Trotz dieser Hindernisse und zusätzlicher Herausforderungen konnte die Datenerhebung für die PISA-Studien in Deutschland erfolgreich abgeschlossen werden. Hierzu waren eine enge Kommunikation, Abstimmung, Flexibilität und ein hoher Einsatz von allen Beteiligten und nicht zuletzt von den Jugendlichen gefragt. Es verdeutlicht, dass die erfolgreiche Durchführung der PISA-Studien in Deutschland nicht nur auf rein technischen Aspekten basiert, sondern auch auf dem Engagement, der Flexibilität und dem Einsatz der beteiligten Personen, die maßgeblich zum Erfolg des Gesamtprojekts ***PISA 2022 in Deutschland*** *beigetragen haben.*

12.1 Eine methodenorientierte Einführung zu PISA 2022

Jörg-Henrik Heine

Die seit dem Jahr 2000 im dreijährigen Zyklus durchgeführte PISA-Studie ist die weltweit größte Studie zur Evaluation der Leistungsfähigkeit der Bildungssysteme in den teilnehmenden Staaten. Als Maß für die Leistungsfähigkeit eines jeweiligen Bildungssystems gelten die Kompetenzen von Jugendlichen in den drei wiederkehrenden Kernbereichen Lesen, Mathematik und Naturwissenschaften. Ergänzt werden diese sogenannten *Kern- oder Hauptdomänen* um innovative Konzepte aus wechselnden Bereichen, wie Problemlösen [*problem solving*] in den PISA-Runden 2003 und 2012 (OECD, 2013), kollaboratives Problemlösen [*collaborative problem solving*] in der PISA-Runde 2015 (OECD, 2017b), globale Kompetenz [*global competence*] in der PISA-Runde 2018 (OECD, 2019) und das kreative Denken [*creative thinking*] in der aktuellen Runde 2022 (OECD, 2023).

Erhoben werden die Kompetenzen fünfzehnjähriger Jugendlicher, die in den meisten Staaten in diesem Alter das frühestmögliche Ende ihrer Pflichtschulzeit erreicht haben. Es muss betont werden, dass bei internationalen Vergleichsstudien wie PISA die Inferenzeinheit zur inhaltlichen Interpretation der erzielten Befunde trotz der Datenerhebung einzelner Jugendlicher auf der Ebene des Systems liegt. So dürfen die Befunde nicht als individuelle Maße für die Performanz einzelner Jugendlicher missverstanden werden, sondern müssen als Maß für Leistungs- und Integrationsfähigkeit des jeweiligen Bildungssystems auch im Hinblick auf Teilpopulationen aufgefasst werden (z. B. Jungen vs. Mädchen, oder Jugendliche mit vs. ohne Zuwanderungshintergrund). Diese übergeordnete Zielsetzung ist auch die substanzwissenschaftliche Begründung für den Einsatz von speziellen *Testdesigns* (vgl. Abschnitt 11.4).

In den letzten zwei Jahrzehnten hat sich PISA in dieser Weise zum weltweiten Maßstab für die vergleichende Messung von Bildungssystemen entwickelt. In mehr als 80 Staaten werden im Rahmen der PISA-Studien regelmäßig Daten erhoben, um die Qualität, Chancengerechtigkeit und die Effizienz der Bildungssysteme zu vergleichen. Die Durchführung einer solch umfangreichen Studie erfordert eine sorgfältige, staatenübergreifende Planung und Organisation bezüglich der angewandten Methodik. Dies gilt sowohl für die Sicherstellung einer vergleichbaren Art der Datenerhebung in den einzelnen Staaten, die Übersetzung und Adaption der Testmaterialien, die Stichprobenziehung, das Datenmanagement, als auch für die Auswertung der Daten im Verbund mit den internationalen Vertragsnehmern. Bei PISA werden daher spezifische Methoden verwendet, die einerseits ein hohes psychometrisch-methodisches Niveau erreichen, andererseits aber auch in Staaten mit den unterschiedlichsten, möglicherweise auch einfachen infrastrukturellen Voraussetzungen realisierbar sein müssen. Übergreifend sind diese Aspekte in den technischen Standards zu der PISA-Erhebung festgehalten (vgl. OECD, 2020a). Die konkreten Abläufe zur Durchführung der einzelnen Schritte für eine

PISA-Erhebungsrunde werden für die nationalen Projektzentren im *Projektmanagement-Manual* für alle teilnehmenden Staaten verbindlich festgehalten (vgl. OECD, 2020b).

Um unter diesen Rahmenvorgaben die Kompetenzen der Jugendlichen als zentralen Indikator für die Leistungsfähigkeit des jeweiligen Bildungssystems vergleichbar zu messen, werden für die zu erfassenden Kompetenzbereiche standardisierte Tests eingesetzt. Die Erhebung relevanter bildungsbezogener Hintergrundinformationen über die Jugendlichen, ihre Eltern und Lehrkräfte sowie ihre Schulen erfolgt ergänzend über Fragebogenverfahren. Ein detaillierter Überblick zu den für PISA (2022) verfügbaren Kompetenzbereichen und Fragebogenmodulen wird in im Abschnitt 12.2 gegeben. Dabei werden zunächst die international *verbindlichen* Kompetenzbereiche und Fragebogenmodule vorgestellt. Zusätzlich werden die durch die OECD angebotenen *optionalen* Kompetenzdomänen und Fragebogenmodule präsentiert und dargelegt, welche davon in Deutschland in der Testung 2022 realisiert wurden.

Sowohl die Aufgaben in den standardisierten Tests als auch die einzelnen Fragen für die Skalen in den Fragebogenmodulen müssen, ausgehend von den in der englischen und französischen Sprache vorliegenden Quellfassungen, in die jeweilige Landesprache übertragen werden. Der aus der reinen *Übersetzung* und gleichzeitig etwaigen *Anpassungen* an kulturelle Besonderheiten bestehender *Übertragungsprozess* erfolgt in den nationalen Projektzentren der jeweiligen Staaten. Dieser mehrstufige Prozess wird für alle teilnehmenden Staaten vom internationalen Projektpartner cApStAn (https://www.capstan.be) begleitet und überwacht, um trotz kultureller Besonderheiten in einzelnen Staaten eine internationale Vergleichbarkeit der eingesetzten Test- und Fragebogeninstrumente sicherzustellen. Seit dem ersten Erhebungszyklus im Jahr 2000 hat cApStAn die Hauptverantwortung für die sprachliche Qualitätssicherung und -kontrolle der übersetzten und angepassten Erhebungsinstrumente übernommen. Im Abschnitt 12.3 in diesem Kapitel werden die einzelnen Schritte des Übertragungsprozesses aus der Perspektive des nationalen Projektzentrums in Deutschland im Verbund mit den beiden deutschsprachigen PISA-Teilnehmerstaaten Österreich und Schweiz dargestellt.

Die übersetzten Testaufgaben sowie die Fragen zu den bildungsbezogenen Hintergrundinformationen werden dann in einer spezifischen Abfolge – dem *Testdesign* – für die Jugendlichen, deren Eltern, Lehrkräfte und Schulen aufbereitet. Seit der PISA-Runde 2015 erfolgt die Darbietung der Testaufgaben und Hintergrundfragebögen für die Jugendlichen in elektronischer Form am Computer. Allerdings besteht (auch in der aktuellen Runde 2022) für einzelne Staaten die Möglichkeit zur Durchführung der Testsitzungen mit papierbasierten Testinstrumenten. So wählten im Jahr 2015 59 von 73 teilnehmenden Staaten und im Jahr 2018 70 von 79 teilnehmenden Staaten die computerbasierte Testung (vgl. OECD, 2017, S. 129-130; OECD, 2021, Tabelle 9.1). Für die aktuelle Erhebung 2022 griffen lediglich vier der teilnehmenden Staaten auf die papierbasierte Testversion zurück (OECD, in Vorb., Tabelle A1.2; s. auch Tabelle 1.1web).

Durch diese deutliche Tendenz zur computerbasierten Testung ergeben sich für das Testdesign vielfältige Möglichkeiten. Im Abschnitt 12.4 des vorliegenden Kapitels wird

auf methodische Aspekte von Testdesigns in internationalen Large-Scale-Assessments wie PISA eingegangen. Dabei werden auch spezifische Punkte zum aktuellen Testdesign bei PISA 2022 sowohl für die Fragebögen als auch für die Kompetenztests behandelt und Implikation für (Sekundär-)Datenanalysen dargestellt und diskutiert.

Eine zentrale Grundlage der Erkenntnisse aus den PISA-Studien bildet die Stichprobe der fünfzehnjährigen Jugendlichen aller teilnehmenden Staaten. Um Rückschlüsse aus der stichprobenbasierten PISA-Erhebung auf die Grundgesamtheit aller Fünfzehnjährigen in den einzelnen Teilnehmerstaaten und damit auf das jeweilige Bildungssystem mit einer hohen Vergleichbarkeit zu ermöglichen, müssen die Verfahren der Stichprobenziehung in allen teilnehmenden Staaten einem streng standardisierten Plan folgen. Die PISA-Studien etablieren dafür ein valides und verbindliches Stichproben-Regelwerk (vgl. z. B. OECD, 2016, 2017a für die PISA-Haupterhebung 2018), welches gewährleistet, dass Stichprobenziehungsstandards in allen teilnehmenden Staaten eingehalten werden. Im Abschnitt 12.5 wird ausführlich auf die Methodik der Stichprobenziehung eingegangen und deren Umsetzung im Jahr 2022 in Deutschland erläutert.

Die praktische Umsetzung von PISA in Deutschland bedarf der Zusammenarbeit unterschiedlicher Personen in den Schulen, der IEA-Hamburg den internationalen Vertragsnehmern der OECD und dem Projektzentrum an der TU-München als gemeinsames Team. PISA in Deutschland berücksichtigt alle Schularten aus allen Bundesländern, daher ist es wichtig, dass die PISA-Testung in allen Schulen und Bundesländern nach den gleichen Standards und Richtlinien abläuft. Besonders herausfordernd war die Durchführung von PISA 2022 in Deutschland während der Coronavirus-Pandemie (SARS-CoV-2). Die Belastungen und Einschränkungen, die mit dem pandemischen Geschehen einhergingen, führten zu zusätzlichen Schwierigkeiten bei der Organisation und Durchführung der Tests. Schulschließungen, Wechselunterricht, Hygienevorschriften und nicht zuletzt krankheitsbedingte Ausfälle erschwerten eine fixe Planung und Umsetzung der Tests, und es waren zusätzliche Maßnahmen erforderlich, um die Sicherheit der Jugendlichen und der anderen in den Schulen beteiligten Personen zu gewährleisten. Im Abschnitt 12.6 werden die organisatorischen Abläufe und die praktische Durchführung der PISA-Erhebungen 2022 im Feld beziehungsweise in den Schulen für Deutschland dargestellt. Daneben werden Aspekte der Datenhaltung im Rahmen des Datenmanagements dokumentiert.

Als Ergebnis der Datenerhebung und des sich daran anschließenden Datenmanagements in Zusammenarbeit mit den internationalen Vertragsnehmern stehen den Projektzentren der teilnehmenden Staaten etwa ab Spätsommer des Folgejahres der Erhebung (für die aktuelle PISA-Runde 2022 ab 5. September 2023) anonymisierte Datensätze für eigene Analysen und Datenauswertungen zur Verfügung. Der Datensatz für die PISA-Kompetenzdomänen enthält zusätzlich zu den kodierten Antworten der Jugendlichen auf die Testaufgaben zehn Schätzer (Plausible Values; PV) für das von ihnen erreichte Kompetenzniveau. Diese Plausible Values erlauben auf der Ebene der beteiligten Staaten die erwartungstreue Schätzung des nationalen Mittelwerts und Streuung in jeder der drei bzw. vier Kompetenzdomänen (vgl. Abschnitt 12.4). Neben den Plausible Values

liegen in den bereitgestellten Datensätzen auch Messwerte für bildungsbezogene Merkmalsausprägungen und Einstellungen, wie sie über die einzelnen Fragen in den Hintergrundfragebögen für die Jugendlichen, deren Eltern und Lehrkräften erhoben werden, vor. Diese Messwerte sind für jedes, über mehrere Fragen operationalisierte Merkmal in den Datensätzen jeweils als fertig skalierte, abgeleitete Variable enthalten – sogenannte *derived variables* (DV; vgl. hierzu auch Abschnitt 12.4). Diese hier beschriebenen PV und DV sind das Ergebnis von Skalierungen durch den internationalen Vertragsnehmer Educational Testing Service (ETS) und bilden die Grundlage für die in den einzelnen Kapiteln des deutschen PISA-Berichtsbandes referierten Ergebnisse.

Für vertiefende Analysen, die in einzelnen Kapiteln berichtet werden, müssen allerdings aus unterschiedlichen Gründen eigene Skalierungen vorgenommen werden, so beispielsweise zur vertiefenden Analyse von Trendentwicklungen in der Unterrichtswahrnehmung (vgl. Kapitel 8) oder zur Untersuchung von spezifischen Fragestellungen zum Lernen unter Pandemiebedingungen (vgl. Kapitel 10). Das methodische Vorgehen bei diesen nationalen Skalierungen und auch die Anwendung von Klassifikationsverfahren zur personenzentrierten Analyse von Mustern der Unterrichtswahrnehmung (vgl. Kapitel 8) werden in diesem PISA-Berichtsband in Abschnitt 12.7 dokumentiert.

Wie in den einzelnen Kapiteln zu den drei Kompetenzdomänen dargestellt (Kapitel 3, 5 und 6), haben sich für die aktuelle PISA-Runde im Rahmen einer vergleichenden Trendbetrachtung der Ergebnisse mit denjenigen aus früheren Runden, teilweise erhebliche Veränderungen (deutliche Rückgänge) im erreichten Kompetenzniveau für Deutschland ergeben. Solche Trendanalysen können von der jeweils angewendeten Skalierungsmethodik und den getroffenen analytischen Entscheidungen bei der Verlinkung einzelner PISA-Runden beeinflusst werden (vgl. z. B. Heine & Robitzsch, 2022; Robitzsch, & Lüdtke, 2021). Im Einzelnen zeigen solche Arbeiten, dass u. a. (1) der Wechsel von papierbasiertem zu computerbasiertem Erhebungsmodus (z. B. Goldhammer et al., 2019; Robitzsch et al., 2016, 2020) sowie (2) die Art der Verlinkung und der Bestimmung der Link-Fehler (z. B. Fischer et al., 2019; Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019) Trendergebnisse beeinflussen können. Eine vergleichende Analyse (vgl. Heine & Robitzsch, 2022) zu unterschiedlichen methodisch-analytischen Entscheidungen bei Large-Scale-Assessments und deren Auswirkungen auf Trendaussagen, sowohl zwischen als auch innerhalb einzelner PISA-Teilnehmerstaaten, zeigt dabei, dass insbesondere die Auswahl der zur Verlinkung verwendeten Aufgaben (Items) und deren unterschiedliche Inhalte einen entscheidenden Einfluss auf das Länderranking und die Entwicklungstrends zwischen und innerhalb der Länder nehmen können (Heine & Robitzsch, 2022). Zur Absicherung der in den einzelnen Kapiteln auf Basis der internationalen Skalierungen berichteten (Trend-)Ergebnisse wurden im Zentrum für internationale Bildungsstudien (ZIB) eigene (marginale) Trendskalierungen, im Sinne einer Kreuzvalidierung, durchgeführt. Die Befunde aus diesen Skalierungen werden in Abschnitt 12.8 berichtet. Die Analysen zeigen, dass beispielsweise für die Trendveränderungen der Mittelwerte zwischen 2018 und 2022 über die drei PISA-Hauptdomänen allenfalls mögliche Verschiebungen im Bereich von rund 0.2 bis 4 PISA-Punkten,

verursacht durch methodische Aspekte sowie Veränderungen in der Zusammensetzung der Population (bedingte Trendschätzung), mitgedacht werden müssten. Insofern kann festgestellt werden, dass im Vergleich zu den Trendaussagen basierend auf den Skalierungen der internationalen OECD-Vertragsnehmer keine grundlegenden Veränderungen in der Kernaussage zu erwarten wären.

Im letzten Abschnitt 12.9 des vorliegenden Kapitels wird aus methodischer Perspektive ein Gesamtfazit zur Durchführung der aktuellen PISA-Studie in Deutschland gezogen.

Literatur

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-Model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling, 61*(1), 37–64.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*(3), 305.

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: Grundbildung im internationalen Vergleich* (S. 163–186). Waxmann.

Heine, J.-H., & Robitzsch, A. (2022). Evaluating the effects of analytical decisions in large-scale assessments: Analyzing PISA mathematics 2003–2012. *Large-scale Assessments in Education*, *10*(1), 10. https://doi.org/10.1186/s40536-022-00129-5

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kroehne, U., Goldhammer, F., & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (S. 383–430). Waxmann. https://www.waxmann.com/buch3555

OECD. (2013). *PISA 2012 Assessment and analytical framework*. OECD Publishing. http://www.oecd-ilibrary.org/education/pisa-2012-assessment-and-analytical-framework_9789264190511-en

OECD. (2016) *Sampling in PISA*. https://www.oecd.org/pisa/pisaproducts/SAMPLING-IN-PISA.pdf

OECD. (2017a) *Main survey school sampling preparation manual*. https://www.oecd.org/pisa/pisaproducts/MAIN-SURVEY-SCHOOL-SAMPLING-PREPARATION-MANUAL.pdf

OECD. (2017b). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2017c). *PISA 2015 Assessment and analytical framework. Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2020a). *PISA 2022 technical standards*. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Technical-Standards.pdf

OECD. (2020b). *PISA national project manager manual*. https://www.oecd.org/pisa/pisaproducts/PISA-2022-National-Project-Manager-NPM-Manual.pdf

OECD. (2021). *PISA 2018 technical report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2023). *PISA 2022 Assessment and analytical framework*. Paris: OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en

OECD. (in Vorb.). *PISA 2022 technical report*. OECD Publishing.

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 444–465. https://doi.org/10.1080/0969594X.2018.1433633

Robitzsch, A., & Lüdtke, O. (2021). Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. *PsyArXiv.* 31 August 2021. https://doi.org/10.31234/osf.io/pkjth

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology, 11,* 884. https://doi.org/10.3389/fpsyg.2020.00884

Robitzsch, A., Lüdtke, O., Köller, O., Kroehne, U., Goldhammer, F., Heine, J.-H., (2016). Herausforderung 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

12.2 Instrumente in PISA 2022
Internationale Pflichten und Optionen sowie nationale Ergänzungen

Jennifer Diedrich

Um die Bildungsergebnisse in Deutschland international vergleichen zu können (und darüber hinaus einen Mehrwert für das nationale Bildungssystem zu erzielen), werden sowohl kognitive Testungen als auch Kontextfragebögen in PISA 2022 eingesetzt. Die Instrumente zur Erfassung kognitiver Merkmale richten sich dabei an die Schüler*innen am Ende ihrer Pflichtschulzeit, während die Kontextfragebögen darüber hinaus die wichtigsten Akteure schulischen Lernens und Lehrens adressieren: Schüler*innen, Lehrkräfte, Schulleitungen und Eltern. Die Instrumente können dabei drei Bereichen zugeordnet werden (s. Tabelle 12.2.1): (1) Instrumente, die international verpflichtend sind (IP), (2) Instrumente, die seitens der internationalen Studienleitung optional angeboten werden (IO) sowie (3) Instrumente, die in Deutschland national ergänzt wurden (ZM & NE). Dieser Abschnitt beschreibt die Instrumente dieser drei Bereiche.

12.2.1 Internationale Pflichtbestandteile

12.2.1.1 Testinstrumente zur Erfassung kognitiver Leistungen

Das Kernelement der PISA-Studie bilden die über eine Testung erfassten Kompetenzbereiche: mathematische Kompetenz (IPK1), naturwissenschaftliche Kompetenz (IPK2; s. Kapitel 5; OECD, 2023g) und Lesekompetenz (IPK3; s. Kapitel 6; OECD, 2023f) sowie die jeweilige innovative Domäne (IPK4). In PISA 2022 bildete Mathematik die Hauptdomäne und wurde somit vertieft erfasst (s. Kapitel 2 und 3; OECD, 2023e sowie Kapitel 4; OECD, 2023a). Die Ergebnisse zur innovativen Domäne – in PISA 2022 kreatives Denken (OECD, 2023b) – werden stets ein halbes Jahr später veröffentlicht und stehen somit noch unter Embargo.

Tabelle 12.2.1: Übersicht der Instrumente in PISA 2022 in Deutschland

	Internationale Pflichtbestandteile (IP)	Internationale Optionen (IO)	Nationale Ergänzungen	
			ZIB-Modul (ZM)	Weitere Ergänzungen (NE)
Testinstrumente zur Testung kognitiver Leistungen (K)	IPK1: Mathematische Kompetenz IPK2: Naturwissenschaftliche Komptenz IPK3: Lesekompetenz IPK4: Kreatives Denken	IOK1(a): Zusatzdomäne Financial Literacy	ZMK1: Berliner Test zur Erfassung fluider und kristalliner Intelligenz ZMK2: Test zum divergenten Denken	
Kontextfragebögen (F)	IPF1: Schüler*innen IPF2: Schulleitungen	IOF1: Eltern IOF2: Lehrkräfte IOF3: ICT-Modul IOF4(b): Well-being-Modul IOF5(b): Global-Competence-Modul IOF6(c): Global-Crises-Module in IPF1 IOF7(c): Global-Crises-Module in IPF2	ZMF3: Zusatzfragen zum schulischen Lehren und Lernen unter Distanzbedingungen für die Schüler*innen	NEF1: National ergänzte Fragen in IPF1 (beginnend immer mit der Nummer ST8**) NEF2: Zusatzfragbogen zum schulischen Lehren und Lernen unter Distanzbedingungen für die Schulleitungen

Anmerkung: (a) = Deutschland nahm in PISA 2022 an dieser Zusatzdomäne nicht teil.
(b) = In Deutschland wurden nur einzelne Fragen aus diesen Modulen umgesetzt.
(c) = Fragebogenmodule wurden erst nach der Verschiebung im Jahr 2020 entwickelt und als internationale Option ergänzt.

12.2.1.2 Kontextfragebögen

Neben den Tests werden Fragebögen eingesetzt, um den Kontext des Lehrens und Lernens zu erfassen (OECD, 2023a). Diese Fragebögen richten sich an Schulleitungen und Schüler*innen. Die hierin erfragten Konstrukte (Abbildung 4.1web) decken seit der ersten PISA-Erhebungsrunde die Felder des klassischen Kontext-Input-Prozess-Outcome-Modells von Purves (1987) ab. Ein Beispiel für eine Kontextfrage wäre die seit der ersten Erhebungsrunde erfragte Unterscheidung zwischen ländlichen und städtischen Schulen (Schulleitungsfragebogen SC001; Mang et al., 2023). Zum Input gehören Fragen wie das Geschlecht der Schüler*innen (Schüler*innenfragebogen ST004; s. auch Abschnitt 1.3.2 in Kapitel 1). Als Prozess werden Konstrukte wie die Offenheit des Schulklimas in Bezug

auf Kreativität (Schulleitungsfragebogen SC208) betrachtet. Den Ergebnissen (Outcomes) wird z. B. die Stressresistenz der Schüler*innen (Schüler*innenfragebogen ST345) zugerechnet. Die Konstrukte decken sowohl domänenspezifische, d. h. auf die jeweilige Haupt- oder innovative Domäne bezogene (z. B. SC208), als auch domänenübergreifende (z. B. ST345) Merkmale ab.

Auf Ebene der Staaten ist die Vorgabe der Fragebögen für Schüler*innen (IPF1) und Schulleitungen (IPF2) obligatorisch. In Deutschland variiert der Verpflichtungsgrad zur Beantwortung der Fragebögen allerdings zwischen den Bundesländern (s. Tabelle 12.2.2).

12.2.2 Internationale Optionen

12.2.2.1 Testinstrumente zur Erfassung kognitiver Leistungen

Seitens der internationalen Studienleitung wurde wie seit 2012 durchgängig als Zusatzdomäne die Financial Literacy (IOK1; OECD, 2023c) angeboten. Dabei wird mit einem Leistungstest sowie mit begleitenden domänenspezifischen Modulen in den Fragebögen die Kompetenz der Schüler*innen im Treffen wohlüberlegter finanzieller Entscheidungen sowie der gezielten Suche nach Unterstützung bei finanziellen Fragen erhoben. Diese internationale Option erfordert eine zusätzliche Stichprobe. Diese rund 1 600 Schüler*innen erhalten lediglich Aufgaben zu Financial Literacy sowie zur mathematischen Kompetenz oder Lesekompetenz. Die gemeinsame Steuerungsgruppe „Feststellung der Leistungsfähigkeit des Bildungswesens im internationalen Vergleich“ aus Bund und Ländern hat - basierend auf einer erneuten Stellungnahme der nationalen Projektleitung von PISA am ZIB - für PISA 2022 entschieden, nicht an dieser internationalen Option (OECD, 2018) teilzunehmen.

12.2.2.2 Kontextfragebögen

Auch in PISA 2022 wurden seitens der internationalen Studienleitung zwei Fragebögen für eine jeweils eigene Zielgruppe (Eltern und Lehrkräfte) sowie vier Fragebogenmodule als Bestandteil der Kontextfragebögen spezifischer Zielgruppen (OECD, 2018; OECD, 2023a) angeboten.

Der *Elternfragebogen* (IOF1) erfragt seit PISA 2006 unter anderem den sozioökonomischen Hintergrund sowie das häusliche (Lern-)Umfeld der Jugendlichen. Dazu erhielten auch in Deutschland in PISA 2022 alle Eltern der getesteten Schüler*innen einen papierbasierten Fragebogen, dessen Beantwortung zuhause stattfand und anonym und freiwillig war. In diesem Berichtsband werden erste Befunde aus dem Elternfragebogen im Kapitel 7 berichtet; zudem werden diese Merkmale bei der Veröffentlichung der Ergebnisse der innovativen Domäne im Frühjahr 2024 eine Rolle spielen.

Tab. 12.2.2: Übersicht der Instrumente in PISA 2022 in Deutschland nach Verpflichtungsgrad in den Bundesländern

	Öffentliche Schulen					Schulen in freier Trägerschaft				
Bundesland	Kompetenztests (IPK1-4)	Schüler*innenfragebogen (IPF1)	Schulleitungsfragebogen (IPF2)	Lehrkräftefragebogen (IOF2)	Elternfragebogen (IOF1)	Kompetenztests (IPK1-4)	Schüler*innenfragebogen (IPF1)	Schulleitungsfragebogen (IPF2)	Lehrkräftefragebogen (IOF2)	Elternfragebogen (IOF1)
Baden-Württemberg	verpflichtend	verpflichtend	freiwillig	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Bayern	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig
Berlin	verpflichtend	verpflichtend	teilverpflichtend	teilverpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Brandenburg	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Bremen	verpflichtend	verpflichtend	teilverpflichtend	teilverpflichtend	freiwillig	verpflichtend	verpflichtend	teilverpflichtend	teilverpflichtend	freiwillig
Hamburg	verpflichtend	freiwillig	teilverpflichtend	teilverpflichtend	freiwillig	verpflichtend	freiwillig	teilverpflichtend	teilverpflichtend	freiwillig
Hessen	verpflichtend	verpflichtend	teilverpflichtend	teilverpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Mecklenburg-Vorpommern	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig
Niedersachsen	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Nordrhein-Westfalen	verpflichtend	freiwillig	verpflichtend	verpflichtend	freiwillig	verpflichtend	freiwillig	verpflichtend	verpflichtend	freiwillig
Rheinland-Pfalz	verpflichtend	freiwillig	teilverpflichtend	teilverpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Saarland	verpflichtend	freiwillig	teilverpflichtend	teilverpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Sachsen	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Sachsen-Anhalt	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig
Schleswig-Holstein	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig	freiwillig
Thüringen	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig	Entscheidung der Schule bzw. des Trägers	Entscheidung der Schule bzw. des Trägers	Entscheidung der Schule bzw. des Trägers	Entscheidung der Schule bzw. des Trägers	freiwillig

Für die Befragung mit dem *Lehrkräftefragebogen* (IOF2) wurden aus jeder an PISA teilnehmenden Schule Lehrkräfte der Hauptdomäne Mathematik sowie weitere Lehrkräfte zufällig ausgewählt (s. Abschnitt 12.5 in diesem technischen Kapitel). Die Befragung fand auf einer geschützten Onlineplattform statt. Der Fragebogen erfasste Konstrukte, die von strukturellen Aspekten, wie dem Stundenumfang, bis hin zu motivational-emotionalen Merkmalen, wie dem Affekt der Lehrkräfte, reichen. Erste Ergebnisse werden in diesem Band in Kapitel 8 berichtet.

Seit der ersten Erhebungsrunde im Jahr 2000 kann der Kontextfragebogen der Schüler*innen um ein Modul *zu Informations- und Kommunikationstechnologien* erweitert werden, in dem Kenntnisse zu und Nutzung von digitalen Medien erfragt werden (ICT; IOF3; OECD, 2023d). Diese Option wurde wie schon in dem vergangenen Erhebungsrunden in Deutschland genutzt. Die entsprechenden Befunde spielen eine zentrale Rolle in Kapitel 9 dieses Berichtsbandes.

An zwei weiteren Fragebogenmodulen, *Well-being* (IOF4) sowie *Global Competence* (IOF5), als internationale Option zum Schüler*innenfragebogen nahm Deutschland nicht in vollem Umfang teil. Stattdessen wurden einzelne Fragen als nationale Ergänzung einbezogen. Details hierzu werden im Skalenhandbuch zu PISA 2022 beschrieben (ZIB, in Vorb.).

Angesichts der Corona-Pandemie und der im Frühjahr 2020 seitens der internationalen Studienleitung entschiedenen Verschiebung des Feldtests und der Haupterhebung um ein Jahr wurden Zusatzmodule zu den Kontextfragebögen der Schulleitungen und der Schüler*innen für PISA 2022 entwickelt, die als *Global-Crises-Module* bezeichnet werden (IOF5 und IOF6; Bertling et al., 2020; OECD, 2023a). Diese Module dienen dazu, das Lehren und Lernen während der pandemiebedingten Einschränkungen des Schulunterrichts aus Sicht der Schulleitungen und der Schüler*innen genauer zu betrachten und wurden durch die Ergänzung entsprechender Fragen in den beiden Pflichtfragebögen (Schüler*innen sowie Schulleitungen) umgesetzt. Die Befunde zu Merkmalen wie der wahrgenommenen familiären Unterstützung während des Distanzunterrichts (ST353), der Teilnahmequote der Schüler*innen am Distanzunterricht (SC220) sowie die Einschätzung der Befragten, wie sich auf einen möglichen zukünftigen Distanzunterricht vorbereitet sehen (ST356 und SC224) werden in Kapitel 10 eingehend beschrieben. Diese Zusatzmodule wurden von der Mehrheit der PISA-Teilnehmerstaaten durchgeführt.

12.2.3 Nationale Ergänzungen

Über die internationalen Pflichtbestandteile sowie internationalen Optionen hinaus war es bei PISA 2022 nötig, nationale Ergänzungen der Instrumente zu entwickeln und vorzugeben. Dies geschah aus drei möglichen Gründen (s. auch Kapitel 1): (1) von internationaler Seite wurden Tests oder Fragebögen nicht fortgeführt, die aber zur Analyse bedeutsamer Trends in Deutschland nötig sind, (2) im Rahmen der internationalen Erfassung fehlen relevante Merkmale oder (3) andere für Bildungspolitik und -wissenschaft wichtige Konstrukte fehlten in den internationalen Instrumenten, sollten aber im nationalen Kontext erhoben werden. In PISA 2022 wurden aus diesen Gründen Ergänzungen an verschiedenen Kontextfragebögen vorgenommen.

Das *ZIB-Modul* ist eine auf dem Itembuilder (Kroehne, 2023) basierte Testanwendung, in der in PISA 2022 drei Module umgesetzt wurden. Erstens wurde als wichtige Kovariate wie schon in PISA 2012 (papierbasiert), 2018 (computerbasiert) die allgemeine kognitive Grundfähigkeit der Schüler*innen mittels des Berliner Tests zur Erfassung fluider und kristalliner Intelligenz (BEFKI; Wilhelm et al., 2014) erfasst (ZMK1). Zweitens wurden zur Validierung des Tests zum kreativen Denken klassische Kreativitätsinstrumente in Form von drei Aufgaben zum divergenten Denken (ZMK2; Diedrich & Lewalter, 2021) national hinzugenommen. Abschließend wurde das ZIB-Modul zur Haupterhebung noch um wenige Fragen zur Teilnahme der Schüler*innen an Förderangeboten im Kontext des pandemiebedingt eingeschränkten Unterrichts ergänzt (ZMF3).

Im allgemeinen *Schüler*innenfragebogen* wurden, wie in bisherigen Erhebungsrunden, spezifisch für Deutschland weitere Merkmale erfragt, deren genaue Aufschlüsselung dem Skalenhandbuch entnommen werden kann (NEF1; ZIB, in Vorbereitung). Diese sind an Itemnummern über 800 erkennbar. Beispielsweise wurden zum familiären Hintergrund ergänzend die EGP-Klassen (s. Kapitel 7) oder für aufschlussreiche Trendanalysen u. a. Freude und Interesse an Mathematik (ST827) erfragt, welches bereits 2003 and 2012 eingesetzt wurde.

Abschließend wurden die *Schulleitungen* gebeten, einen zweiten Fragebogen über einen gesonderten Link online auszufüllen (NEF2). Darin wurde über die Fragen aus dem Global-Crises-Modul hinaus spezifisch für Deutschland wichtige Fragen zur Zeit des Distanzunterrichts, vor allem aber der Umsetzung von Förderangeboten im Nachgang des Distanz- bzw. pandemiebedingt eingeschränkten Unterrichts vorgelegt (s. Kapitel 10).

Literatur

Bertling, J., Rojas, N., Alegre, J., & Faherty, K. (2020). *A tool to caputre learning experiences during Covid-19: The PISA global crises questionnaire module*. OECD Publishing. https://doi.org/10.1787/9988df4e-en

Diedrich, J., & Lewalter, D. (2021, October 29). *Validierung Kreatives Denken in PISA 2022: Stellungnahme des ZIB zur innovativen Domäne in PISA 2022: Kreatives Denken.* (unpublished)

Kroehne, U. (2023). *Open computer-based assessment with the CBA ItemBuilder*. https://cba.itembuilder.de

Mang, J., Seidl, L., Schiepe-Tiska, A., Tupac-Yupanqui, A., Ziernwald, L., Doroganova, A., Weis, M., Diedrich, J., Heine, J.-H., Gonzalez-Rodriguez. E. & Reiss, K. (2023). *PISA 2018 Skalenhandbuch. Dokumentation der Erhebungsinstrumente* (2., aktualisierte Auflage). Waxmann. https://doi.org/10.31244/9783830994961

OECD. (2018). *PISA 2021 international options*. (unpublished)

OECD. (2023a). PISA 2022 Context questionnaire framework: Balancing trends and innovation. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 169–237). OECD. https://doi.org/10.1787/471ae22e-en

OECD. (2023b). PISA 2022 creative thinking framework. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 140–168). OECD Publishing. https://doi.org/10.1787/471ae22e-en

OECD. (2023c). PISA 2022 financial literacy framework. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 99–139). OECD Publishing. https://doi.org/10.1787/471ae22e-en

OECD. (2023d). PISA 2022 ICT framework. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 238-285). OECD Publishing. https://doi.org/10.1787/471ae22e-en

OECD. (2023e). PISA 2022 mathematics framework. In OECD (Hrsg.), In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 18–98). OECD Publishing. https://doi.org/10.1787/471ae22e-en

OECD. (2023f). PISA reading framework. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 286). OECD Publishing. https://doi.org/10.1787/471ae22e-en

OECD. (2023g). PISA science framework. In OECD (Hrsg.), *PISA 2022 assessment and analytical framework* (S. 287). OECD Publishing. https://doi.org/10.1787/471ae22e-en

Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review, 31*(1), 10–28. https://doi.org/10.1086/446653

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI 8-10)*. Hogrefe. https://www.testzentrale.de/shop/berliner-test-zur-erfassung-fluider-und-kristalliner-intelligenz-fuer-die-8-bis-10-jahrgangsstufe.html

ZIB. (2025, in Vorb.). *Skalenhandbuch PISA 2022*. Waxmann.

12.3 Der Übersetzungs- und Anpassungsprozess der Testinstrumente

Elisabeth Gonzalez Rodriguez, Julia Mang & Jörg-Henrik Heine

Die Datenerhebungen für PISA werden in den teilnehmenden Staaten mit unterschiedlichen Sprachen, Kulturen und Schulsystemen durchgeführt. Für eine vergleichbare Messung der PISA-Kompetenzdomänen und zur Erfassung der Informationen über den Hintergrund und das Bildungsumfeld der Jugendlichen werden standardisierte Tests und Hintergrundfragebögen eingesetzt. Die Äquivalenz der in die jeweilige Testsprache übertragenen nationalen Versionen ist eine zentrale Voraussetzung für international vergleichbare Daten. Dies wird über strenge Qualitätsstandards für die Übersetzung der Testmaterialien erreicht, die beispielweise in den *translation and adaptation guidelines* [dt.: Übersetzungs- und Adaptionsrichtlinien] festgehalten sind (vgl. cApStAn & Halleux, 2019a). Die Einhaltung dieser Richtlinien wird von dem im Jahre 2000 gegründeten internationalen Vertragsnehmer cApStAn (*The **c**ompany of **Ap**or, **St**eve and **An**drea*; vgl. https://www.capstan.be/history/) staatenübergreifend begleitet und überwacht. Für den Übersetzungs- und Anpassungsprozess der Testmaterialien stellt cApStAn den nationalen Projektzentren unterschiedliche Materialien und (Online-)Tools zur Verfügung (vgl. z. B. cApStAn & Halleux, 2019b; cApStAn 2019, 2022). Der Übersetzungs- und Anpassungsprozess der Testmaterialien erstreckt sich über einen Zeitraum von insgesamt etwa 14 Monaten (mit Unterbrechungen) und beinhaltet dabei mehrere Einzelschritte, bei denen die Übersetzungen und kulturellen Anpassungen zwischen den Übersetzer*innen, dem nationalen Projektzentrum und dem internationalen Vertragsnehmer cApStAn mehrfach ausgetauscht und wechselseitig überprüft werden. Dieser mehrstufige Prozess ist von entscheidender Bedeutung dafür, dass keine (systematischen) Verzerrungen entstehen, welche negative Auswirkungen auf die internationale Vergleichbarkeit der PISA-Ergebnisse haben könnten.

Insbesondere wird dabei auf folgende Aspekte geachtet: Allgemein soll (1) die Verständlichkeit von Texten, Grafiken und Tabellen, die in den einzelnen Fragen eingesetzt werden, im Sinne des verwendeten sprachlichen Niveaus, durch die Übersetzung und Anpassung nicht verändert werden; wobei hier (2) bei den Testaufgaben zu den Kompetenzdomänen auf einzelne sprachliche Formulierungen geachtet wird, die möglicherweise die Art der kognitiven Prozesse und / oder die notwendigen Lösungsstrategien für einzelne Testaufgaben verändern könnten. Für den Bereich der Hintergrundfragebögen wird (3) darauf geachtet, dass Unklarheiten aufgrund von unterschiedlichen kulturellen Hintergründen in den Frageformulierungen vermieden werden.

Insgesamt lässt sich der Übersetzungsprozess für die Testaufgaben und für die Formulierungen der Hintergrundfragebögen in drei Bereiche aufteilen:

1. *Double Translation* (doppelte Übersetzung)
2. *Reconciliation* (Abgleich und Zusammenführen der Übersetzungen)
3. *Verification* (Verifizierung der Übersetzung)

Der reine Übersetzungsprozess der Testaufgaben und Fragebögen in die deutsche Sprache dauert ca. 4 Wochen. Dafür erstellen zwei Übersetzer*innen zwei individuelle und unabhängige Übersetzungsvorschläge. Die Übersetzer*innen müssen eine staatlich geprüfte Ausbildung haben sowie durch das Landgericht vereidigt sein. Wenn möglich sollten sie bereits Erfahrung aus vorherigen PISA Übersetzungsprozessen haben. Eine Vertraulichkeitsvereinbarung mit dem nationalen Projektzentrum (National Center) sichert die Geheimhaltung der entwickelten Testaufgaben. Die Originaltexte liegen in den vorgegebenen Sprachen Englisch und Französisch vor (die Amtssprachen der OECD). Die Fachübersetzung erfolgt dann Wort für Wort ins Deutsche, wobei Standards für die Übersetzung wiederkehrender Begriffe und sprachlicher Wendungen sowie bestimmte (Fach-)Begriffe berücksichtigt werden.

Im nächsten Schritt werden die beiden Übersetzungsvorschläge im *Reconciliation*-Schritt gegeneinander abgeglichen und daraus eine gemeinsame Version erstellt. Dieser Schritt nimmt in etwa 8 Wochen in Anspruch. Dieser Abgleich der Übersetzungen erfolgt im National Center in Rücksprache und Kooperation mit ausgewählten Fachdidaktiker*innen und Expert*innen (wie zum Beispiel Lehrkräfte) sowie im DACHLVerband, der die Staaten Deutschland, Österreich, Schweiz[1], und Luxemburg[2] umfasst. Nationale Anpassungen, wie zum Beispiel Fragen zu deutschen Schularten oder deutschen Schul- und Ausbildungsabschlüssen, werden ebenfalls während dieser Phase eingearbeitet. Am Ende liegt eine sogenannte „*Common German Version*" aller deutschsprachigen Staaten, welche an PISA teilnehmen, vor.

Diese Common German Version wird nun in mehreren Schleifen im Austausch mit dem nationalen Projektzentrum verifiziert. Der Verifier, also der internationale PISA-Vertragsnehmer cApStAn aus Belgien gibt hierzu Rückmeldungen und macht gegebenenfalls. Anpassungsvorschläge. Danach folgen zusätzliche Sichtungen durch weitere internationale PISA-Vertragsnehmer, die technische Kontrollen und Layout-Überprüfungen der finalen Testinstrumente vornehmen sowie einen letzten Check durch die Testentwickler*innen koordinieren. In einem letzten Schritt wird die fertiggestellte Version im nationalen Projektzentrum abschließend geprüft, bevor die Testinstrumente vor dem Einsatz im PISA-Feldtest zur datenschutzrechtlichen Überprüfung in den Bundesländern freigegeben werden. Die für den Feldtest erarbeiteten Übersetzungen werden so auch in der Haupterhebung eingesetzt.

1 Die Zusammenarbeit bei den Übersetzungen mit Schweiz beziehen sich auf den deutschsprachigen Teil der Schweiz.

2 Luxemburg hat an der PISA-Runde 2022 nicht teilgenommen, weswegen der Übersetzungs- und Anpassungsprozess ohne Beteiligung Luxemburgs erfolgte.

Literatur

cApStAn. (2019). *PISA 2022: Instructions for Field Trial PVS (Preferred Verification Schedule)*. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Instructions-for-Field-Trial-PVS-Preferred-Verification-Schedule.pdf

cApStAn. (2022). *OECD PISA | Translation Validation | cApStAn*. https://www.capstan.be/casestudies/oecd-pisa/

cApStAn, & Halleux, B. (2019a). *PISA 2022 translation and adaptation guidelines* (ETS and Core A Contractors, Hrsg.). https://www.oecd.org/pisa/pisaproducts/PISA-2022-Translation-and-Adaptation-Guidelines.pdf

cApStAn, & Halleux, B. (2019b). *PISA 2022 translation kit* (ETS and Core A Contractors, Hrsg.). https://www.oecd.org/pisa/pisaproducts/PISA-2022-Translation-and-Adaptation-Guidelines.pdf

12.4 Testdesign und Populationsmodell in PISA

Jörg-Henrik Heine

Mit dem Begriff *Testdesign* wird – in Abgrenzung zum allgemeinen (experimentellen) Design einer wissenschaftlichen Untersuchung – im Folgenden die besondere Art und Weise der sequenziellen Anordnung einzelner Testaufgaben für die Kompetenzdomänen und die Fragen (Items) in den Hintergrundfragebögen und deren Zuordnung zu den Jugendlichen im Rahmen der PISA-Erhebung verstanden.

Das grundlegende Prinzip des Testdesigns für die Kompetenzdomänen in PISA basierte auf einer Variante des *Matrixsamplings* (vgl. Rutkowski, 2013). Beim Matrixsampling werden nicht nur die zu testenden Jugendlichen, sondern auch die Aufgaben (Items) nach einem zufälligen Prinzip ausgewählt. Jeder einzelnen Person in der Stichprobe wird dabei nur eine Teilmenge von Aufgaben aus dem gesamten Aufgabenpool zu Bearbeitung vorgelegt. Insgesamt bearbeiten so verschiedene Gruppen von Jugendlichen unterschiedliche, sich jedoch insgesamt überschneidende Mengen von Testaufgaben. Das Prinzip des Matrixsamplings orientiert sich an der Zielsetzung von vergleichenden Bildungsstudien (vgl. auch Abschnitt 12.1 zu diesem Kapitel), wonach die Inferenzeinheit der Ergebnisse auf der Systemebene (Staatenebene) liegt und nicht auf der Individualebene, die auf einzelne Jugendliche bezogen wäre. Das Ziel der Messung von Kompetenzen in PISA ist also die statistische Beschreibung von Populationen und Teilpopulationen des jeweiligen Bildungssystems (vgl. z. B. von Davier et al., 2006).

Die dem PISA Testdesign zugrundeliegende zentrale Idee, nicht nur die Auswahl der Jugendlichen, sondern auch die der Testaufgaben (Items) stichprobenbasiert vorzunehmen, geht zurück auf frühe Überlegungen zu der in der bildungswissenschaftlichen Testpraxis typischen Situation, dass die Testzeit einerseits eng begrenzt ist und andererseits ein breites Spektrum an Kompetenzbereichen mit einer angemessenen Anzahl von Aufgaben (Items) valide und reliabel gemessen werden soll (vgl. Johnson & Lord, 1958; Lord, 1962).

12.4.1 Unvollständige Versuchspläne und Booklet-Designs

Das Prinzip einer zufälligen Auswahl von Items und deren balancierte Zuweisung zu den Testpersonen über verschiedenen Testformen hinweg ist eine bekannte Methode aus dem Bereich der experimentellen Versuchsplanung (vgl. Federer & Nguyen, 2002). Experimentelle Versuchspläne werden in (natur-)wissenschaftlichen Untersuchungen eingesetzt, um unerwünschte Zusammenhänge zwischen dem zu messenden Merkmal und bestimmten experimentellen (Rahmen-)Bedingungen zu kontrollieren (Fisher & Yates, 1963). Typischerweise ist es dabei in der Regel nicht möglich, sämtliche experimentellen Bedingungen zu berücksichtigen (bei PISA die Kombination von Personen und Aufgaben), weswegen unvollständige, aber dennoch balancierte Versuchspläne realisiert

werden, die als *balanced incomplete block designs* (BIBD) [dt. balancierte unvollständige Block-Designs] bezeichnet werden (z. B. Banerjee, 1948; Yates, 1936; Youden, 1937, 1962).

Die Zusammenstellung des PISA-Testdesigns nach diesen Prinzipien bezieht sich allerdings nicht auf jede einzelne Aufgabe der Kompetenzdomänen. Vielmehr werden die einzelnen Aufgaben (Items) zunächst nach inhaltlichen Kriterien in sogenannten *Item-Clustern* gruppiert, welche sodann in balancierter Form zu einem Testdesign zusammengestellt werden. Die Gesamtheit des auf diese Weise aus mehreren Versionen bestehenden Testmaterials wir oft auch als „*Rotiertes-Booklet-Design*" bezeichnet.

Weitere Informationen zu dieser Methodik und ihre praktische Umsetzung in PISA und anderen internationalen Vergleichsstudien finden sich in den Arbeiten von Rutkowski et al. (2013) und Weeks et al. (2013) sowie – in Bezug auf das aktuelle PISA-Testdesign – bei Yamamoto et al. (2018a, 2018b sowie Yamamoto et al., 2019).

Mit dem Einsatz solcher Booklet-Designs ist – im Vergleich zu Testdesigns, in denen jede teilnehmende Person dieselbe Menge von Aufgaben erhält (beispielsweise Schulaufgaben) – eine zentrale Implikation bei der Testauswertung verbunden. So ist es allein durch die Verteilung der Aufgaben auf unterschiedliche Testformen nicht länger sinnvoll, bei der Testauswertung Statistiken zu verwenden, die auf der Anzahl der richtig gegebenen Antworten basieren. Unterschiede in der Gesamtpunktzahl oder in darauf basierenden Statistiken zwischen Jugendlichen, die verschiedene Testformen (in den Booklets) bearbeitet haben, könnten beispielsweise allein auf unterschiedliche Schwierigkeitsgrade der Booklets und der darin enthaltenen Aufgaben zurückzuführen sein. Diese und vergleichbare Einschränkungen in der Testauswertung werden durch die Anwendung von Skalierungsmodellen aus der Item-Response-Theorie (IRT) überwunden (vgl. Berezner & Adams, 2017; Heine et al., 2016). Durch die IRT-Skalierung können sowohl die Kompetenzen der Jugendlichen als auch die Schwierigkeiten der Testaufgaben (Items) auf einer gemeinsamen Skala abgetragen werden, dies gilt auch dann, wenn nicht alle Jugendliche identische Mengen von Items bearbeiten haben. Auf diese Weise können die Kompetenzverteilungen in unterschiedlichen Populationen (den einzelnen Staaten) oder in Teilpopulation beschreiben werden und die Beziehungen zwischen dem Kompetenzniveau und unterschiedlichen Merkmalsausprägungen oder zu Einstellungen aus den Hintergrundvariablen für die (Teil-)Population geschätzt werden.

12.4.2 Item- und Test-Information und Messgenauigkeit

Der psychometrische Begriff *test targeting* beschreibt im Rahmen der Item-Response-Theory (IRT) das Verhältnis der Verteilung der Itemschwierigkeiten zu der Kompetenzverteilung der getesteten Personen. Ein optimales *test targeting* liegt dann vor, wenn sich zentrale Verteilungsparameter (Mittelwert und Standardabweichung) für die Itemschwierigkeiten (σ) und die gemessenen Kompetenzen (θ) angleichen. Bezogen auf einzelne zu testende Personen bedeutet dies, dass die Information aus der Messung dann am

genauesten ausfällt, wenn die Kompetenz θ_v der Person *v* etwa der Schwierigkeit σ_i einer Aufgabe *i* entspricht. Formal lässt sich dieses Prinzip einer möglichst optimalen Passung von Personenkompetenz und Aufgabenschwierigkeit anhand des psychometrischen Konzeptes der *Item-* bzw. *Testinformation* im Rahmen der IRT darstellen. Es zeigt sich dabei, dass in IRT-Modellen die Messpräzision eines Items als Funktion des Kontinuums der zu messenden Merkmalsdimension (θ) variiert, was durch die Item-Informationsfunktion dargestellt werden kann (vgl. Abbildung 12.4.1). Aus der IRT-Modellgleichung, beispielsweise für das Rasch-Modell (vgl. Rasch, 1960) für dichotome Item-Antworten, ergeben sich zunächst die Antwortkategoriewahrscheinlichkeiten $p(X_i = x_i)$ mit $x_i \in \{0,1\}$ für die beiden Antwortkategorien „*falsch*" = 0 und „*richtig*" = 1, gegeben den Vektor der Messwerte für die Personenkompetenz (θ) mit der Itemschwierigkeit σ_i einer einzelnen Testaufgabe *i* (vgl. Gleichung 1).

$$p\left(X_i = x_i|\theta, \sigma_i\right) = \frac{e^{(x_i \cdot (\theta - \sigma_i))}}{1 + e^{(\theta - \sigma_i)}}; x_i \in \{0, 1\} \qquad (1)$$

Die Steigung des Graphs der Funktion (der Item Characteristic Curve – ICC) der Lösungswahrscheinlichkeiten $p(X_i = 1)$ für eine richtige Antwort, hat ihren größten Wert gerade dann, wenn die Itemschwierigkeit σ_i mit der Kompetenzausprägung (θ) übereinstimmt (vgl. Abbildung 12.4.1 hellblaue Linie). Vergleicht man beispielsweise zwei Personen mit unterschiedlicher Kompetenzausprägung (θ) anhand eines Items *i*, so sind deutliche Unterschiede in den Lösungswahrscheinlichkeiten (y-Achse) nur dann zu erwarten, wenn die Kompetenzausprägungen (θ) im Bereich der Itemschwierigkeit σ_i liegen. Die Steigung der Funktion der Lösungswahrscheinlichkeiten $p(X_i = 1)$ gibt also an, wie hoch der Gewinn an Information durch Anwendung des Items *i* bei einer Person mit einer bestimmten Kompetenzausprägung (θ) ist und wird als *Item-Informationsfunktion* bezeichnet. Aus der Modellgleichung des Rasch-Modells (vgl. Gleichung 1) lässt sich die Lösungswahrscheinlihckeit $p(X_i = 1)$ für eine richtige Antwort wie folgt schreiben (vgl. Gleichung 2).

$$p\left(X_i = 1|\theta, \sigma_i\right) = \frac{e^{(1 \cdot (\theta - \sigma_i))}}{1 + e^{(\theta - \sigma_i)}} \qquad (2)$$

Die Steigung einer Funktion, wie in Gleichung 2, ist allgemein durch deren erste Ableitung gegeben. Die erste partielle Ableitung nach θ von Gleichung (2) ergibt sich nach Anwendung der Quotientenregel und Vereinfachung der Summe im Zähler wie folgt in Gleichung 3.

$$\frac{\partial}{\partial\theta} p\left(x_i = 1|\theta, \sigma_i\right) = \frac{e^{(\theta - \sigma_i)}}{\left(1 + e^{(\theta - \sigma_i)}\right)^2} \qquad (3)$$

Man kann zeigen (vgl. Fischer, 1974, S. 295), dass sich die numerische Ausprägung I_i der Item-Informationsfunktion $I_i(\theta)$ für ein Item *i* bei einer bestimmten, gegebenen Kompetenzausprägung θ als Produkt aus der Lösungswahrscheinlichkeit $p(X_i = 1$; „*richtig*“) und deren Gegenwahrscheinlichkeit 1- $p(X_i = 0$; „*falsch*“) ergibt (vgl. Gleichung 4).

$$I_i|\theta = p\,(X_i = 1|\theta) \cdot p\,(X_i = 0|\theta) \qquad (4)$$

Betrachtet man ein Kontinuum für die Differenzen aus θ *und* σ_i auf der Logit-Metrik des Rasch-Modells mit einem typischen Wertebereich von *-3 bis +3,* so ergibt sich der in Abbildung 12.4.1 mit dunkelblauem Linienzug gezeigte Graph der Item-Informationsfunktion $I_i(\theta)$.

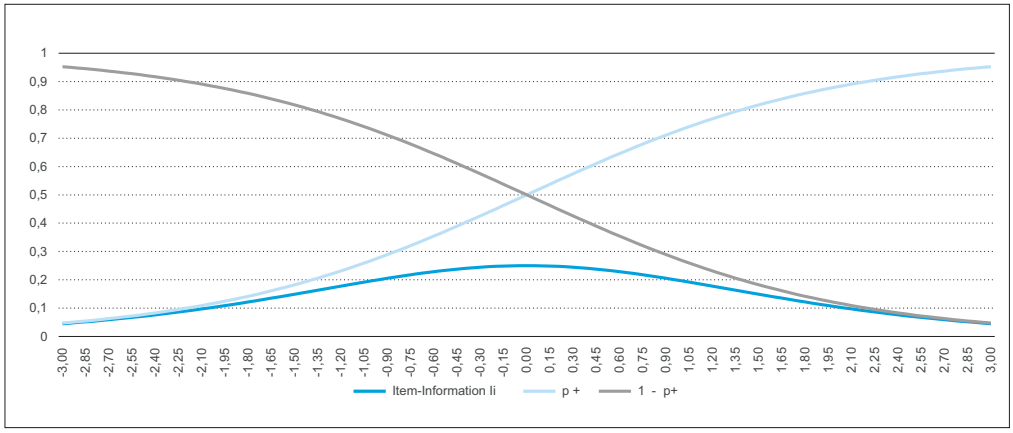


Abbildung 12.4.1: Graph der Item-Informationsfunktion und der Antwortkategoriewahrscheinlichkeiten im Rasch-Modell für dichotome Antwortskalen.

Wie Abbildung 12.4.1 zeigt, erreicht der Funktionsgraph der Item-Information (dunkelblaue Linie) seinen Maximalwert für die Item-Information mit I_i = 0.25 an der Stelle (auf der x-Achse), an der die Differenz aus θ und σ_i null ist. An derselben Position ist gleichzeitig die Lösungswahrscheinlichkeit (richtige Antwort) für das Item 0.5 (vgl. hellbaue und graue Linie in Abbildung 12.4.1). Praktisch gesprochen bedeutet dies, dass zu testende Personen zum Erreichen einer maximalen Messgenauigkeit idealerweise nur Aufgaben bearbeiten sollten, bei denen die Wahrscheinlichkeit, dass sie diese Aufgaben richtig lösen, 0.5 beträgt.

Für einen aus mehreren Aufgaben (Items) bestehenden Test lässt sich – bei Modellgeltung – für jede beliebige Person *v* mit dem Personenparameter θ_v durch Addition der zuvor berechneten einzelnen Item-Informationsbeträge $I_{i,v}$ die sogenannte Testinformation *I* für den gesamten Test berechnen.

Die zentralen Punkte dieser Betrachtungen bestehen darin, dass sie formal zeigen, dass die Item- und Test-Information und damit die Messgenauigkeit eines Tests mit fest ausgewählten Aufgaben (1) interindividuell (je nach Kompetenzausprägung) variiert und (2) dass die Messgenauigkeit in (extremen) unteren und oberen Kompetenzbereichen eher geringer ausfällt. Bei international vergleichenden Studien wie PISA variiert die Kompetenz beispielsweise in den drei Kerndomänen sowohl innerhalb einzelner Staaten, aber auch zwischen den Staaten erheblich (z. B. Rutkowski et al., 2019) – insbesondere auch durch die zunehmende Teilnahme von immer mehr Staaten (vgl. Kamens & McNeely, 2010). Diesem Umstand ist bei der (Weiter-)Entwicklung eines Testdesigns für PISA Rechnung zu tragen. Es muss an dieser Stelle angemerkt werden, dass die bei PISA tatsächlich zur Skalierung der Antworten der Jugendlichen zu den Testaufgaben eingesetzten IRT-Modelle formal wesentlich komplexer ausfallen, als es hier am Beispiel der Modellgleichung des Rasch-Modells gezeigt wurde. Dennoch gelten die am einfachen Beispiel skizzierten Prinzipien der Item- und Test-Information sowie des *test targeting* auch für Modellerweiterungen (vgl. Kubinger, 2016, S. 109) wie multidimensionale IRT-Modelle (Adams et al.,1997) oder Zwei-Parameter Modelle (z. B. Muraki, 1992), welche auf dem Rasch-Modell aufbauen und jeweils die Grundlage der Skalierung bei PISA bilden. Die psychometrischen Grundlagen der eingesetzten IRT-Skalierungsmodelle werden für die PISA-Runden 2000 bis 2012 bei Adams et al. (1997) beschrieben; für die Runden seit PISA 2015 sei auf die Beiträge von Khorramdel et al. (2019), von Davier (2005; 2017) sowie auf den Technischen Report zu PISA 2015 (OECD, 2017) verwiesen.

Um insgesamt die Messungenauigkeiten auf der Ebene einzelner Personen angemessen zu berücksichtigen, verwendet PISA seit der ersten Runde im Jahr 2000 Fähigkeitsschätzer aus multiplen Imputationen, die als *Plausible Values* (dt. plausible Werte) bezeichnet werden. Im Gegensatz zu einfachen Punktschätzern für jeden Jugendlichen in der Stichprobe (im Sinne eines festen Zahlenwerts als Maß für das erreichte Kompetenzniveau) werden in jeder der drei Kompetenzdomänen zehn *Plausible Values* gezogen (vgl. z. B. von Davier et al., 2009). Die grundlegende Idee besteht darin, zunächst individuelle a-posteriori-Verteilungen der erfassten Kompetenzdomänen auf Basis der beobachteten Antwortdaten unter Hinzunahme von Informationen aus dem Fragebogen für die Jugendlichen als latentes Regressionsmodell mit Modellen der Item-Response-Theory zu modellieren. Der Vorteil dieses *Populationsmodells* besteht darin, dass Statistiken, wie zum Beispiel Mittelwerte für einzelnen Staaten und Mittelwertunterschiede zwischen Staaten, unter Berücksichtigung der Messfehler auf Personenebene unverzerrt bestimmt werden können.

12.4.3 Computerbasierte Testung und adaptive Testdesigns

In einem computerbasierten Testdesign bestehen gegenüber festen, papierbasierten Testdesigns wesentlich größere Freiheitsgrade in der Art und Weise, wie die einzelnen Aufgabencluster den Jugendlichen nach multiplen Kriterien in optimaler Balancierung, gerade auch nach Aufgabenschwierigkeit, zugewiesen werden können. So wies bereits das erste computerbasierte Testdesign für PISA 2015 insgesamt 2 376 theoretisch mögliche, individuelle Testzusammenstellungen auf (vgl. Heine et al., 2016, Anhang B), welche auf 66 Basistestformen, kombiniert mit 36 Clusterkombinationen für jeweils zwei naturwissenschaftliche Aufgabencluster basierten.

Für die letzte Erhebungsrunde im Jahre 2018 wurde das Prinzip einer flexiblen und individualisierten Form der Testzusammenstellung am Computer noch weiter ausgebaut. Das neue Element des Testdesigns bestand dabei – damals nur für den Kerninhaltsbereich Lesen – in der Einführung einer *adaptiven Komponente* als zusätzliches Kriterium zur Anordnung der Aufgabencluster in den so computerbasiert individualisierten Testformen. Die *adaptive Komponente* besteht dabei in einer dynamisch-systematischen Zuweisung von Aufgabenclustern (Testlets) zu einzelnen Jugendlichen anhand ihrer – in vorher bearbeiteten Aufgaben gezeigten – Kompetenzniveaus.

Für PISA 2018 umfasste die adaptive Komponente drei Stufen, die lediglich für den Kerninhaltbereich der Lesekompetenz angewendet wurde (vgl. Heine & Reiss, 2019). Einen technisch detaillierten Überblick zur Einführung des mehrstufig adaptiven Testdesigns seit der PISA-Runde 2018 geben Yamamoto et al. (2019). Für die aktuelle PISA-Runde 2022 wurde das Prinzip des adaptiven Testdesigns für die Lesekompetenz und den Kerninhaltsbereich der Mathematik angewendet (vgl. OECD, 2018).

Aus messtheoretischer Perspektive ist mit dem Einsatz des adaptiven Testens zunächst ganz elementar die Möglichkeit zur Verkürzung der Testzeit verbunden (z. B. Kubinger, 2016). Dieser Zeitgewinn bei der Datenerhebung kann genutzt werden, um andere Konstrukte zu erheben. Allerdings besteht das Ziel der adaptiven Tests bei PISA nicht unbedingt (nur) darin, einzelne Tests oder die Testung zu verkürzen. Vielmehr sollen über ein mit der Adaptivität verbundenes besseres *test targeting* auf Personen- und Staatenebene die Unterschiede in der Messgenauigkeit auf den PISA-Skalen verringert werden. Praktisch soll mit der Weiterentwicklung fester Booklet-Designs, wie sie auch für eine papierbasierte Testung eingesetzt werden konnten, hin zu adaptiven (mehrstufigen) Testdesigns die Messgenauigkeit im Bereich der niedrigeren und hohen Kompetenzniveaus verbessert werden. Mit der Einführung und dem Ausbau des adaptiven Testens kann also eher die Erwartung einer (weiteren) Reduktion und vor allem aber einer Homogenisierung des Messfehlers entlang der (internationalen) Kompetenzverteilung verknüpft werden – ohne dabei die Testzeit für die Jugendlichen zu erhöhen (Oranje et al., 2014). Testtheoretisch kann diese Annahme plausibel bereits durch das – durch die Adaptivität erzielte – bessere *test targeting* für jede einzelne Person in der Stichprobe begründet werden (vgl. auch Abbildung 12.4.1). Allerdings muss hierzu einschränkend angemerkt werden, dass sich die adaptive Komponente des Testdesigns bei PISA in den

Runden 2018 und 2022 nicht auf einzelne Aufgaben (Items) sondern auf ganze Item-Cluster (vgl. Abschnitt 12.4.1) bezieht (Yamamoto et al., 2018b). Ferner erfolgt daher die adaptive Zuweisung nicht nach der Item- oder Testinformation (vgl. Abschnitt 12.4.2) sondern nach der Anzahl gelöster Aufgaben in einzelnen *Testlets* (vgl. Abschnitt 12.4.4.1) in einer der adaptiven Stufen (vgl. Heine & Reiss, 2019). In dieser praktischen Umsetzung als dreistufiges adaptives Testdesign fallen die theoretisch möglichen Gewinne in der Messpräzision daher nicht maximal aus (vgl. Yamamoto et al., 2018b; 2018c), weswegen die Argumentation hinsichtlich einer Erhöhung der Messgenauigkeit durch die adaptive Komponente auch kritisch gesehen werden kann (vgl. Robitzsch & Lüdtke, 2021).

12.4.4 Implikationen aus computerbasierten adaptiven Testdesigns

Die zunehmende Verwendung eines mehrstufig adaptiven Testdesigns bei PISA hat verschiedene Implikationen, und zwar aus einer methodischen Perspektive sowohl im Hinblick auf die Gestaltung des adaptiven Testdesigns selbst als auch auf die anschließende Skalierung und Auswertung der Daten. Daneben ergeben sich aus einer computerbasierten Testvorgabe auch positive Aspekte einer vereinfachten Testadministration im Feld sowie die Möglichkeit der gleichzeitigen Erfassung und späteren Analyse von zusätzlichen Log-Daten, die während der Testdurchführung vom Computersystem aufgezeichnet werden. Ferner wird aus einer inhaltlichen, psychologischen Perspektive seit den frühen Anfängen des adaptiven Testens dessen Auswirkungen auf die Testmotivation diskutiert. In den folgenden Abschnitten werden diese Implikationen einer computerbasierten und adaptiven Testvorgabe bei PISA skizziert.

12.4.4.1 Aufgaben-Routing in adaptiven Testdesigns und Skalierung

Mit der Entwicklung eines mehrstufig-adaptiven Testdesigns geht unmittelbar die Frage einher, auf welcher Grundlage die Zuordnung der getesteten Personen zu Testaufgaben einer folgenden Stufe des adaptiven Designs erfolgen soll. In vollständig IRT-basierten adaptiven Testdesigns erfolgt die Auswahl der jeweils folgenden Aufgabe auf Basis einer kontinuierlichen Bestimmung der Personenparameter („Kompetenz“) jeweils nach der Bearbeitung einer Aufgabe (z. B. Wainer & Dorans, 2000). Das *Routing* der zu testenden Personen zu den jeweils nächsten Aufgaben kann dabei beispielsweise über das Kriterium der Item-Information (wie oben dargestellt) erfolgen. Das Prinzip besteht darin, möglichst viel diagnostische Information über die individuelle Kompetenzausprägung mit möglichst wenigen Aufgaben zu erlangen. Die Umsetzung dieses Prinzip setzt ein unmittelbares *scoring* (Bewertung) voraus – also beispielsweise bei einem dichotomen Antwortformat die unmittelbare Bewertung der gegebenen Antworten nach den beiden Kategorien 1 = richtig oder 0 = falsch. Allerdings benötigen bei PISA rund ein Drittel

aller Aufgaben eine Bewertung (Kodierung) durch menschliche Kodierer – sogenannte *human coded* Items (vgl. Yamamoto et al., 2018c). Dies führt dazu, dass bei PISA für das Routing im mehrstufig adaptiven Testdesign nur bestimmte Aufgaben eingesetzt werden können, bei denen ein unmittelbares *scoring* möglich ist – typischerweise handelt es sich dabei um Aufgaben mit fest vorgegeben Antwortkategorien. Für die Kodierung der Testaufgaben in PISA bestehen aktuelle Entwicklungsziele daher auch darin, ein sogenanntes *machine-supported coding system* zum Beispiel perspektivisch auch für Testaufgaben mit offen formuliertem Antwortformat zu entwickeln (z. B. Yamamoto et al., 2018a).

In den PISA-Runden 2018 und 2022 bestand das mehrstufig adaptive Testdesign (*multi-stage adaptive test design*; MSAT) aus drei Stufen, wobei für jede Stufe mehrere, zu sogenannten *Testlets* (vgl. Lewis & Sheehan, 1990) zusammengestellte Aufgabengruppen bereitgestellt werden, die für die adaptiven Stufen 2 und 3 als einfache und schwierige Testlets zusammengestellt wurden (vgl. Heine & Reiss, 2019) für eine detaillierte Darstellung des Designs). Die Zuordnung (*Routing*) der Jugendlichen zu den beiden adaptiven Teststufen erfolgt bei der PISA-Testung derzeit (2022 wie auch 2018) im Grundsatz einfach nach der Anzahl korrekter Antworten in einer der jeweils vorausgehenden (adaptiven) Stufen. Dabei können nur diejenigen Aufgaben berücksichtig werden, die sich aufgrund eines geschlossenen Antwortformats direkt automatisch kodieren lassen.

Allerdings erfolgt die Zuweisung der Jugendlichen zu einem schwierigen oder einfachen Testlets nicht ausschließlich nach der in der bereits bearbeiteten adaptiven Teststufe gezeigten Kompetenz. Stattdessen wird die rein kompetenzorientierte Aufgabenzuweisung zusätzlich von einem probabilistischen Auswahlprinzip überlagert (eine *probability layer,* vgl. Yamamoto et al., 2018c), sodass aus den Gruppen der hoch oder niedrig kompetenten Jugendlichen jeweils etwa (nur) 90 Prozent den folgenden Aufgaben ausschließlich nach ihrem erreichten Kompetenzniveau zugewiesen werden. Dieses Prinzip soll sicherstellen, dass für sämtliche Aufgaben eine Mindestmenge an Aufgabenbearbeitungen vorliegen, was eine wichtige Voraussetzung für die Skalierung im Rahmen der IRT darstellt. Die variierende Anzahl der Aufgabenbearbeitungen aufgrund eines adaptiven Testdesigns wird in der Literatur auch unter dem Begriff der *Item-Exposition* diskutiert (vgl. z. B. Stocking & Swanson, 1993).

Insgesamt können bei der Festlegung der Rahmenbedingungen eines adaptiven (*Testlet*-basierten) Testdesigns also prinzipiell mehrere analytische Entscheidungen a priori getroffen werden (vgl. Luo & Kim, 2018). Im Wesentlichen betreffen diese Entscheidungen Fragen nach den angewendeten Kriterien für das Routing (adaptive Zuweisung der Jugendlichen zu den Aufgaben), zu Mechanismen zur Kontrolle der Item-Expositionsraten sowie zur Länge (Anzahl der Aufgaben) innerhalb eines einzelnen Testlets. Svetina et al. (2019) untersuchen in einer Simulationsstudie die Effekte von solchen Entscheidungen zu globalen Randbedingungen eines mehrstufig adaptiven Testletdesigns auf die erzielten Item-Expositionsraten und die resultierende Genauigkeit der IRT-Parameterschätzung. Insgesamt variierten Svetina et al. (2019) drei zentrale Faktoren des Testdesigns: die Anzahl der Aufgaben pro adaptiver Teststufe, die Methode des Routings sowie

als dritter Faktor des experimentellen Designs der Simulationsstudie den Einsatz eines probabilistischen Routing-Prinzips. Bei diesem Prinzip werden die Jugendlichen mit einer bestimmten Wahrscheinlichkeit (ungleich null) unabhängig von den bisher beantworteten Aufgaben zu der nächsten Teststufe weitergeleitet, was so (mit unterschiedlicher Wahrscheinlichkeit) zu einer entweder optimalen oder suboptimalen Passung zwischen Kompetenz und Aufgabenschwierigkeit in der nächsten adaptiven Teststufe führen kann. Dieses zunächst kontraintuitiv erscheinende Prinzip, welches auch im aktuellen PISA-Testdesign angewendet wird (vgl. Yamamoto et al., 2019) soll sicherstellen, dass auch ein Teil der leistungsstarken Jugendlichen mit leichten Aufgaben konfrontiert wird, um etwaige Boden- oder Deckeneffekte (vgl. Rutkowski et al., 2019) aufgrund von variierenden Item-Expositionsraten bei der IRT-Schätzung der Aufgabenschwierigkeiten zu kontrollieren.

Svetina et al. (2019) schlussfolgern, basierend auf den Ergebnissen, zunächst übergreifend, dass die eindeutige Identifikation einer idealen Faktorstufenkombination (analytische Entscheidungen zu Randbedingungen des Testdesigns) für ein adaptives Testdesign durch die Simulation nicht erreicht werden konnte. Allerdings zeigen sich bei isolierter Betrachtung der einzelnen Faktoren Tendenzen zu deren Auswirkungen auf die IRT-Parameterschätzung und die Item-Expositionsraten. So gelingt die Personenparameterschätzung am besten, wenn das Routing IRT-basiert erfolgt, entweder basierend auf der Testinformation (Fisher-Information) oder basierend auf vorläufigen Personenparameterschätzungen. Gleichzeitig fallen aber – wenig überraschend – die Item-Expositionsraten homogener aus, wenn der Anteil eines suboptimalen (unabhängig von der vorläufig gemessenen Kompetenz) Routing-Prinzips steigt, was wiederum Auswirkungen auf die Schätzbarkeit beziehungsweise auf die Schätzgenauigkeit der IRT-Modellparameter nehmen kann. Gerade die Gegenüberstellung dieser beiden zentralen Befunde aus der Simulationsstudie verdeutlichen, dass die Bewertung der Angemessenheit eines adaptiven, Testlet-basierten Testdesigns im Grunde von der Gewichtung einzelner Kriterien abhängt. Für eine optimale Parameterschätzung und -identifikation im Rahmen des anschließend eingesetzten IRT-Skalierungsmodells müssen beispielsweise die Item-Expositionsraten möglichst homogen ausfallen.

Um bei der Skalierung der Kompetenzdomänen in PISA eine möglichst optimale Parameterschätzung zu erzielen, werden seit den ersten Runden multidimensionale Item-Response-Theory (MIRT)-Modelle eingesetzt (z. B. das *Multidimensional Random Coefficients Multinomial Logit Model* (MCLML), vgl. Adams et al., 1997). Obwohl also bereits ein MIRT-Modell eingesetzt wurde, bei dem jedes Item eindeutig nur einer Kompetenzdimension (latente Variable) zugeordnet ist, werden aus praktischen Gründen häufig zusätzlich separate unidimensionale Skalierungen vorgenommen. Die Skalierungen erfolgen dabei für jede Kompetenzdomäne (latente Variable) einzeln mit entsprechenden eindimensionalen IRT-Modellen. Neuere Untersuchungen – beispielsweise von Jewsbury und van Rijn (2020) – zeigen aber, dass bei einer getrennten unidimensionalen Skalierung von Daten aus adaptiven Designs negative Auswirkungen auf die Schätzung der Modellparameter (z. B. die Kompetenzen der Jugendlichen) beobachtet werden

können. Dies liege daran, so argumentieren Jewsbury und van Rijn (2020), dass einige Items, die bei der Routing-Entscheidung verwendet wurden, von den separaten unidimensionalen Skalierungsmodellen nicht berücksichtigt werden. Folgt man den Befunden von Jewsbury und van Rijn (2020), so würde dies bedeuten, dass mit dem Einsatz von adaptiven Testdesigns bei PISA unmittelbar die ausschließliche Anwendung von multidimensionalen IRT-Skalierungsmodellen verbunden wäre.

12.4.4.2 Testadministration und Log-Daten

Neben den oben skizzierten messtheoretischen Vorteilen sind mit dem Einsatz einer computerbasierten Testung auch praktische, administrative Vorteile verbunden. So werden seit 2015 sowohl die Tests zur Erfassung der Kompetenzen als auch die Fragebögen für die Jugendlichen computerbasiert vorgegeben. Im Vergleich zu den in vergangenen Runden zwischen 2000 und 2012 eingesetzten papierbasierten Instrumenten ermöglicht dieser Ansatz eine effiziente und ökonomische Datenerhebung. Während der Testsitzungen mit den Jugendlichen wird dazu ein computerbasiertes System eingesetzt, das über einen USB-Stick entweder mit den vorhandenen Schulcomputern oder mit mitgebrachten tragbaren Computern verbunden wird. Dieses sogenannte *Student Delivery System* (SDS) befindet sich als serverähnliche Softwarestruktur auf den USB-Sticks und benötigt für die Benutzung durch die Jugendlichen keinerlei Internetverbindung. Die Antworten der Jugendlichen werden dabei direkt auf die USB-Sticks in kennwortgeschützte komprimierte Dateien gespeichert. Durch den Einsatz der computerbasierten Testdurchführung entfällt der aufwendige Druck papierbasierter Fragebögen und Testhefte. Diese Testhefte mussten jeweils in mehreren Varianten (Rotationen) produziert werden und konnten nur mit einem beträchtlichen logistischen Aufwand in den einzelnen Schulen termingerecht zur Verfügung gestellt werden.

Ein weiterer Vorteil der computerbasierten Erhebung mit dem SDS besteht in der Möglichkeit einer zusätzlichen Erfassung sogenannter Log-Event-Daten. Log-Event-Daten enthalten Meta- beziehungsweise Para-Informationen (vgl. Kroehne & Goldhammer, 2018) zur Testsituation und zur Art und Weise, wie die Personen in der Testsituation auf das vorgegebene Testmaterial reagieren. Log-Event-Daten können insbesondere im Kontext der PISA-Studien eine wichtige Datengrundlage zusätzliche Auswertungen darstellen (Becker et al., 2022). Durch die Analyse von Log-Event-Daten in Kombination mit den Antwortdaten können auch Erkenntnisse zur Nutzung von elektronischen Lern- und Assessment-Plattformen gewonnen werden (z. B. Goldhammer et al., 2021). Ferner ist die Verwendung von Log-Event-Daten nach der American Educational Research Association (2014) eine relevante Quelle für Validitätsnachweise der eingesetzten Testinstrumente. Die Auswertung von solchen Daten hat zudem in der Lehr-Lern-Forschung bereits zu wichtigen Erkenntnissen geführt. Eine Studie von Stadler et al. (2019) nutzte beispielsweise Log-Event-Daten, um den erfolgreichen bzw. erfolglosen Strategieeinsatz bei der Lösung komplexer Probleme zu analysieren. Im gleichen Sinne argumen-

tieren Goldhammer und Zehner (2017), dass die Verwendung von Log-Event-Daten in Kompetenztests wichtige Hinweise zu meta-kognitiven Strategien sowie zu affektiven Zustände bieten.

12.4.4.3 Psychologisch-motivationale Effekte durch adaptives Testen

Seit Beginn des adaptiven Testens in den 1970er Jahren (vgl. Lord, 1971) sind damit neben einer Effizienzsteigerung auch Erwartungen zu psychologischen Effekten verbunden. So werden beispielsweise eine Verringerung der Testangst (z. B. Akhtar et al., 2023; Rocklin & O'Donnell, 1987) und die Erwartung einer motivationssteigernden Wirkung bei der Testbearbeitung (z. B. Linacre, 2000; Mead & Drasgow, 1993; Wainer & Dorans, 2000; Wise, 2014;) diskutiert. Begründungen für eine motivationssteigernde Wirkung adaptiven Testens konzentrierte sich dabei im Wesentlichen auf einen tatsächlich gegebenen Effekt, der mit Testfrustration zusammenhängt: Durch die dynamisch angepasste Aufgabenauswahl wird vermieden, dass weniger kompetenten Personen zu schwierige Items zur Bearbeitung vorgegeben werden, was potenziell Frustration auslöst. Oft wird auch als theoretische Begründung für die positiven Effekte einer adaptiven im Vergleich zu einer nicht adaptiven Testvorgabe auf die Testmotivation das entwicklungspsychologische Konzept der *Zone der proximalen Entwicklung* (Vygotsky, 1962) angeführt (vgl. z. B. Asseburg, 2011), wonach eine ideale Lernsituation dann besteht, wenn ein sich entwickelnder Mensch einerseits eine hinreichend große Herausforderung zu meistern hat, die ihn andererseits aber weder unter- noch überfordert. Allerdings muss hierzu angemerkt werden, dass Vygotskys Konzept zunächst wenig Aussagen zu motivationalen Effekten macht und sich auch nicht auf eine allein zu bewältigende formale Prüfungs- oder Testsituation bezieht. Vielmehr geht es bei Vygotskys Konzept um die Differenz zwischen dem tatsächlichen Entwicklungsstand, welcher durch *eigenständiges* Problemlösen ermittelt werden kann, und dem potenziellen Entwicklungsstand, welcher durch *gemeinsames* Problemlösen, beispielsweise unter Anleitung von Erwachsenen oder in Zusammenarbeit mit kompetenteren Gleichaltrigen bestimmt werden kann. So zitiert Chaiklin (2003) Vygotsky zur Zone der proximalen Entwicklung wie folgt „*what the child is able to do in collaboration today he will be able to do independently tomorrow*" [*was das Kind heute in Zusammenarbeit tun kann, wird es morgen selbständig tun können*] (Vygotsky, 1987, S. 211, zitiert nach Chaiklin, 2003, S. 40).

Einen weiteren zentralen Aspekt zur Testmotivation stellen die durch die zu testenden Personen wahrgenommenen Konsequenzen aus der Testung dar. Gerade in sogenannten *Low-Stakes*-Testsituationen [dt. etwa: Testsituationen mit niedrigem Einsatz], die dadurch gekennzeichnet sind, dass das Abschneiden im Test keinerlei persönliche Konsequenzen für die Testperson hat (vgl. Wise & DeMars, 2005), kann die Art und Weise der Aufgabenvorgabe einen entscheidenden Einfluss haben (z. B. Wise, 2014). Während in der sogenannten *High-Stakes*-Testsituation [dt. etwa: Testsituationen mit hohem Einsatz] in der Regel nahezu alle Testpersonen allein durch die wahrgenommene

hohe Relevanz der Test- oder Prüfungsergebnisse hoch motiviert sind (z. B. Smith & Smith, 2002; Sundre, 1999), kann die Motivation in *Low-Stakes*-Testsituationen wie PISA (vgl. Akyol et al., 2021) eingeschränkt sein. Insbesondere von leistungsstarken Jugendlichen kann dabei ein adaptives Testszenario, in dem stets Aufgaben ausgewählt werden, die dem gerade gemessenen Kompetenzniveau entsprechen und somit nur mit einer mittleren Wahrscheinlichkeit von 0.5 gelöst werden können, auch demotivierend empfunden werden. Der wahrgenommene Umstand, dass in einem so gestalteten Testlauf unabhängig von der eigenen Anstrengung im Durchschnitt nur die Hälfte der vorgelegten Fragen richtig beantwortet werden kann, entspricht nicht dem üblichen, eigenen Kompetenzerfahrungen von leistungsstaken Jugendlichen. Im Gegensatz dazu könnten leistungsstarke Personen bei nicht adaptiven Tests typischerweise einen deutlich größeren Anteil der Fragen beantworten, was im Vergleich zum adaptiven Testlauf zu einer höheren Motivation führen kann. So zeigen beispielsweise Frey et al. (2009) empirisch, dass die Motivation zur Testbearbeitung bei einer adaptiven Testvorgabe deutlich niedriger ausfällt als bei einer nicht adaptiven Testvorgabe der ansonsten gleichen Aufgabeninhalte. Demgegenüber deuten Befunde aus einer Studie von Ling et al. (2017) darauf hin, dass der Einfluss der Form der Testvorgabe im Kontext eines mathematischen Leistungstests eher geringe Auswirkungen hat. Allerdings wurden in dieser Studie unterschiedlich gestaltete adaptive Testversionen eingesetzt. Eine adaptive Testversion, bei der bewusst leichtere Aufgaben vorgegeben wurden – also Aufgaben mit einer individuellen Lösungswahrscheinlichkeit die über $p = 0.5$ liegt –, führte dabei zu höherem Engagement und geringerer Testängstlichkeit im Vergleich zu der nicht vereinfachten adaptiven Testversion oder im Vergleich zu einer Testversion mit fester Aufgabenvorgabe. Darüber hinaus zeigte sich in allen drei Arten der Testvorgabe eine Leistungszunahme, wenn die Jugendlichen eine sofortige Rückmeldung erhielten (Ling et al., 2017). Eine neuere Meta-Analyse von Akhtar et al. (2023) zu der Frage nach den Effekten adaptiven Testens auf die Testmotivation kommt zu dem Schluss, dass insgesamt die konkret realisierten Testbedingungen bei der adaptiven Testung differenzielle Effekte auf die Testmotivation haben können.

12.4.5 Testdesigns für PISA-Fragebögen

Die psychometrische Grundlage bei der Erfassung der Merkmale und Einstellungen der an den PISA-Erhebungen teilnehmenden Personen stellt (wie auch bei den Kompetenzdomänen) das Konzept der *latenten Variable* dar, deren indirekter empirischer Bezug (vgl. Heine & Reiss, 2019) im Rahmen einer *reflektiven* Operationalisierung durch Fragen in den Fragebogen messbar wird (vgl. z. B. Heine, 2020).

Für den Fragebogenbereich sind hier aber aus messtheoretischer Perspektive zwei Arten von Operationalisierungen zu unterscheiden. Einerseits werden, wie auch für die Kompetenzbereiche, latente Variablen angenommen, wie zum Beispiel Ängstlichkeit in Bezug auf Mathematik oder Motivation, welche typischerweise reflektiv operationalisiert

sind. Andererseits werden zusammengesetzte Indikatoren genutzt, die eine *formative* Operationalisierung als Grundlage haben.

Es gibt zentrale konzeptionelle Unterschiede zwischen den beiden messtheoretischen Ansätzen. Bei den formativ zusammengesetzten Indikatoren basiert die in der Regel angewendete Aufsummierung zur Verrechnung einzelner Fragen (Antworten der Personen) zu einem Index auf einer starken theoretischen Grundlage. Dabei werden einzelne, notwendige Inhaltsbereiche als sogenannte *Facetten* des zu messenden Konzeptes fest definiert, um damit den Index zu formen (vgl. Borg & Mohler, 1993). Beispiele dafür sind in der Soziologie und Ökonometrie weit verbreitete Indizes wie zum Beispiel der bei PISA verwendete Index zum sozioökonomischen Hintergrund (ESCS, vgl. Avvisati, 2020).

Die reflektive Operationalisierung hingegen sieht einzelne, konkret realisierte Fragen (Items) in einer Fragebogenskala lediglich als eine (zufällige), möglichst repräsentative Auswahl an, welche aus einem Item-Universum mit mehreren möglichen Items stammen (vgl. Borg, 1992). Einzelne Fragen werden dabei nur als eine mögliche Auswahl an manifesten Indikatoren angesehen, in denen sich die zu messende latente Variable so zu sagen reflektiert (vgl. Bühner, 2021, S. 16-23).

In der aktuellen PISA-Erhebung (2022) wird für die reflektiven Indikatoren in den Fragebögen für die Jugendlichen ein sogenanntes *within construct rotated design* [dt. etwa: Frage-Rotation innerhalb des Konstruktes] angewendet (vgl. Bertling et al., 2019; OECD, 2023). Analog zu den Kompetenztests wird die Item-Auswahl zur Messung eines bestimmten Konstruktes für einzelne Gruppen von Jugendlichen variiert und systematisch rotiert. Analog zu dem Testdesign für die drei Kompetenzdomänen bedeutet dies, dass für eine reflektiv operationalisierte latente Variable (z. B. Ängstlichkeit in Mathematik; ANXMAT) ein Pool von Fragen als manifeste Indikatoren bereitgestellt wird, von denen allerdings jedem einzelnen Jugendlichen aus der PISA-Stichprobe in den Fragebögen jeweils nur eine Teilmenge aus dem gesamten Fragenpool zu Beantwortung vorgelegt wird. Insgesamt bearbeiten so verschiedene Gruppen von Jugendlichen unterschiedliche, sich jedoch insgesamt überschneidende Mengen von Fragebogen-Items – allerdings eben jeweils für alle Skalen (latenten Variablen), die in den Hintergrundfragbögen erfasst werden.

Demgegenüber wurde noch in der PISA-Runde 2012 eine sogenannte *Skalen-Rotation* angewendet (vgl. Adams, 2013; OECD, 2014). Bei der Skalen-Rotation wurden jeweils ganze Skalen auf insgesamt drei unterschiedliche papierbasierte Fragebögen-Booklets verteilt, sodass (je nach zugewiesenem Booklet) manchen Gruppen von Jugendlichen ganze Inhaltsbereiche mit den entsprechenden Skalen nicht vorgelegt wurden. Dieses noch bei PISA 2012 verfolgte Prinzip führte bei einer gleichzeitigen Analyse unterschiedlicher Skalen (Konstruktbereiche) aus dem Hintergrundfragbogen dazu, dass teilweise nur mit zwei Dritteln bzw. nur mit einem Drittel der Gesamtstichprobe gearbeitet werden konnte (vgl. z. B. Lazarides et al., 2022) – oder auch, dass Analysen mit einer bestimmten Kombination von Skalen nicht möglich sind. Solche, für die Arbeit mit PISA-Daten im Rahmen von Sekundäranalysen nachteilige Restriktionen werden

mit dem *within construct rotated design* in PISA 2022 umgangen. Die Umsetzung dieses neuen *within construct rotated design* für die Skalen in den Hintergrundfragebögen wird durch die Verwendung eines computerbasierten Testdesigns und -systems erleichtert. Ein weiterer Vorteil der computerbasierten Datenerhebung für die Hintergrundfragebögen und die Kompetenztests besteht darin, dass neben den Antwortdaten auch Log-Daten aufgezeichnet werden (z. B. Kroehne & Goldhammer, 2018). Gerade für den Bereich der Hintergrundfragebögen stellen solche Daten zur individuellen Art und Weise der Fragebogenbearbeitung eine informative Datenquelle dar (Becker et al., 2022), welche die aus den Antworten gewonnenen (Skalen-)Messwerte zu bildungsbezogenen Merkmalen und Einstellungen ergänzen können.

Die Messwerte aus den so administrierten Skalen in den Hintergrundfragebögen sind in den PISA-Datensätzen jeweils als fertig skalierte, abgeleitete Variablen enthalten – sogenannte *derived variables* (DV). Im Gegensatz zu den Kompetenzdomänen werden hier allerdings Punktschätzer als Messwerte für die jeweiligen Merkmalsausprägungen und Einstellungen (latente Variable) bereitgestellt. Diese Punktschätzer sind als sogenannte *weighted likelihood estimates* (WLE – Warm, 1989) als Personenparameter ebenfalls das Ergebnis einer IRT-Skalierung durch den internationalen Vertragsnehmer Educational Testing Service (ETS) und bilden die zentrale Grundlage für die in den einzelnen Kapiteln des deutschen PISA-Berichtsbandes referierten Ergebnisse.

12.4.6 Fazit zu Testdesigns für PISA

In diesem Abschnitt der Methodendokumentation zu PISA wurden aktuelle Aspekte des eingesetzten Testdesigns vorgestellt. Das computerbasierte Testdesign bezieht sich dabei auf die in PISA erhobenen Kompetenzdimensionen und auch auf die Hintergrundfragebögen für die Jugendlichen. Die von der OECD getroffene Entscheidung zum Einsatz einer *adaptiven* Komponente im computerbasierten Testdesign für die Kompetenzdomänen muss als Folge aus den Zielsetzungen bei der Weiterentwicklung der PISA-Studien gewertet werden. Insgesamt bietet die computerbasierte und adaptive Datenerhebung einige Vorteile gegenüber einer papierbasierten Datenerhebung. Im Hinblick auf die Kompetenzmessung sind diese Vorteile im Wesentlichen (1) eine zu erwartende Homogenisierung und potenziell gesteigerte Messgenauigkeit über das internationale Kontinuum der Kompetenzdomänen hinweg, (2) die Möglichkeit zur Erhebung von Log- oder Prozessdaten (auch Paradaten, vgl. Kroehne & Goldhammer, 2018) sowie (3) eine effizientere Testdurchführung (vgl. zusammenfassend auch Heine & Reiss, 2019). Die ebenfalls in diesem Abschnitt dargestellten, aus der aktuellen Literatur bekannten, Implikationen aus einer adaptiven Testung im Hinblick auf die Testmotivation indizieren zukünftige vertiefende Analysen zur PISA-Methodik. Gerade für die Daten aus der aktuellen Erhebung 2022 bietet sich hier beispielsweise eine vergleichende Analyse zwischen einer adaptiven Testadministration und einer linearen Testvorgabe für den Kompetenzbereich Mathematik an. So wurde bei der Kompetenztestung für den Bereich Mathe-

matik einer zufällig ausgewählten Teilmenge der Jugendlichen in der PISA-Stichprobe eine fixe, lineare und somit nicht adaptive Testform vorgelegt. Aus diesen Daten ergibt sich für künftige methodisch orientierte Analysen die Möglichkeit, die Auswirkungen der Testvorgabe (adaptiv vs. nicht adaptiv) einerseits auf die erzielten Kompetenzniveaus und andererseits auch auf die Testmotivation im internationalen Vergleich vertiefend zu untersuchen.

Literatur

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education*, *1*, 5. https://doi.org/10.1186/2196-0739-1-5

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial Logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The effect of computerized adaptive testing on motivation and anxiety: A systematic review and meta-analysis. *Assessment*, *30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research*, *42*(2), 184–243. https://doi.org/10.1007/s12122-021-09317-8

American Educational Research Association (Hrsg.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests*. Christian-Albrechts Universität.

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-scale Assessments in Education*, *8*(1), 8. https://doi.org/10.1186/s40536-020-00086-x

Banerjee, K. S. (1948). Weighing designs and balanced incomplete blocks. *The Annals of Mathematical Statistics*, *19*(3), 394–399. https://doi.org/10.1214/aoms/1177730204

Becker, B., Neuendorf, C., & Jansen, M. (2022). Nutzung von Logdaten in der empirischen Bildungsforschung. Eine Bedarfsanalyse. *KonsortSWD Working Paper*, *3*, 1–19. https://doi.org/10.5281/ZENODO.7030996

Berezner, A., & Adams, R. J. (2017). *Why large-scale assessments use scaling and Item Response Theory*. In P. Lietz, J. Cresswell, K. Rust, & R. J. Adams (Hrsg.), *Implementation of large-scale education assessments* (S. 92–136). John Wiley. https://doi.org/10.1002/9781118762462.ch13

Bertling, J. P., ETS, Core B2, OECD, & Alegre. (2019). *PISA 2021 context questionnaire framework*. https://www.oecd.org/pisa/sitedocument/PISA-2021-questionnaire-framework.pdf

Borg, I. (1992). *Grundlagen und Ergebnisse der Facettentheorie*. Huber.

Borg, I., & Mohler, P. Ph. (1993). Zur Indexbildung in der Facettentheorie. *ZUMA-Nachrichten*, *17*(33), 10–24.

Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4., korrigierte und erweiterte Auflage). Pearson.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V. S. Ageyev, & S. M. Miller (Hrsg.), *Vygotsky's educational theory in cultural context,* (S. 39-64). Cambridge University Press. https://doi.org/10.1017/CBO9780511840975.004

Federer, W. T., & Nguyen, N.-K. (2002). Incomplete block designs. In A. H. El-Shaarawi & W. W. Piegorsch (Hrsg.), *Encyclopedia of environmetrics* (S. 1039–1042). John Wiley.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* Huber.

Fisher, R. A., & Yates, J. F. (1963). *Statistical tables for biological, agricultural and medical research*. Oliver and Boyd.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20.

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*(1), 20. https://doi.org/10.1186/s40536-021-00113-5

Heine, J.-H. (2020). *Untersuchungen zum Antwortverhalten und zu Modellen der Skalierung bei der Messung psychologischer Konstrukte* [Monographie, Universität der Bundeswehr]. https://athene-forschung.unibw.de/132861

Heine, J.-H., & Reiss, K. (2019). PISA 2018 – die Methodologie. In K. Reiss, M. Weis, E. Klieme, und O. Köller (Hrsg.), *PISA 2018 Grundbildung im internationalen Vergleich* (S. 241–58). Waxmann.

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kröhne, U., Goldhammer, F., & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Hrsg.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (S. 383–430). Waxmann. https://www.waxmann.com/buch3555

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics*, *45*(4), 383–402. https://doi.org/10.3102/1076998619881790

Johnson, M. C., & Lord, F. M. (1958). An empirical study of the stability of a group mean in relation to the distribution of test items among students. *Educational and Psychological Measurement*, *18*(2), 325–329. https://doi.org/10.1177/001316445801800209

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54(*1), 5–25. https://doi.org/10.1086/648578

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y.-S. Lee (Hrsg.), *Handbook of diagnostic classification models: Models and model Extensions, applications, software packages* (S. 603–628). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_30

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kubinger, K. D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (Hrsg.), *Principles and methods of test construction: standards and recent advances* (S. 104–119). Hogrefe.

Lazarides, R., Schiepe-Tiska, A., Heine, J.-H., & Buchholz, J. (2022). Expectancy-value profiles in math: How are student-perceived teaching behaviors related to motivational transitions? *Learning and Individual Differences*, *98*, 102198. https://doi.org/10.1016/j.lindif.2022.102198

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(4), 367–386. https://doi.org/10.1177/014662169001400404

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come.* MESA Psychometric Laboratory.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259–267. https://doi.org/10.1177/001316446202200202

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*(3), 227–242. https://doi.org/10.1007/BF02297844

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243–263. https://doi.org/10.1111/jedm.12174

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. A*pplied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

OECD. (2014). *PISA 2012 technical report*. OECD Publishing. http://www.oecd-ilibrary.org/education/pisa-2009-technical-report_9789264167872-en

OECD. (2017). *PISA 2015 technical report*. OECD Publishing.

OECD. (2018). *PISA 2022 integrated design*. OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Integrated-Design.pdf

OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group-score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Hrsg.), *Computerized multistage testing: Theory and applications* (S. 371–390). Chapman and Hall/CRC. https://doi.org/10.1201/b16858

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks pædagogiske Institut.

Robitzsch, A., & Lüdtke, O. (2021). Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. *PsyArXiv.* 31 August 2021. https://doi.org/10.31234/osf.io/pkjth

Rocklin, T. R., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*(3), 315–319. https://doi.org/10.1037/0022-0663.79.3.315

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2013). Assessment design for 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. 75–95). CRC Press. https://doi.org/10.1201/b16061

Rutkowski, L., Rutkowski, D., & Liaw, Y.-L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy & Practice*, *26*(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

Smith, L. F., & Smith, J. K. (2002). Relation of test-specific motivation and anxiety to test performance. *Psychological Reports*, *91*(3, Pt1), 1011–1021. https://doi.org/10.2466/PR0.91.7.1011-1021

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00777

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292. https://doi.org/10.1177/014662169301700308

Sundre, D. L. (1999). Does examinee motivation moderate the relationship between test consequences and test performance? *Annual Meeting of the American Educational Research Association*.

Svetina, D., Liaw, Y.-L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in international large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192–213. https://doi.org/10.1111/jedm.12206

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2017). New results on an improved parallel EM algorithm for estimating generalized latent variable models. In L. A. Van Der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Hrsg.), *Quantitative Psychology: The 81st annual meeting of the Psychometric Society, Asheville, North Carolina, 2016* (Bd. 196). Springer International Publishing. https://doi.org/10.1007/978-3-319-56294-0

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI monograph series: Issues and methodologies in large-scale assessments, 2*, 9–36.

von Davier, M., Sinharay, S., & Oranje, A. (2006). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics 26: Psychometrics* (S. 1039–1055). Elsevier. https://doi.org/10.1016/S0169-7161(06)26032-2

Vygotsky, L. (1962). *Thought and language*. MIT Press. https://doi.org/10.1037/11193-000

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2. Aufl.). Lawrence Erlbaum. https://doi.org/10.4324/9781410605931

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, (3), 427–450.

Weeks, Davier, M. von, & Yamamoto, K. (2013). Design considerations for the programme for international student assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Hrsg.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (S. 259-275) CRC Press.

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(1–4), 1–17.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Yamamoto, K., He, Q., Shin, H. J., & von Davier, M. (2018a). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling*, *60*(2), 145–164.

Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018b). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling*, *60*(3), 347–368.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018c). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, *37*(4), 16–27. https://doi.org/10.1111/emip.12226

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018*. OECD. https://doi.org/10.1787/b9435d4b-en

Yates, F. (1936). Incomplete randomized blocks. *Annals of Eugenics,* 7(2), 121–40. https://doi.org/10.1111/j.1469-1809.1936.tb02134.x

Youden, W. J. (1937). Use of incomplete block replications in estimating Tobacco-Mosaic virus. *Contributions from Boyce Thompson Institute, 9*(1), 41–48.

Youden, W. J. (1962). *Experimentation and measurement*. National Science Teachers Association. https://doi.org/10.1119/1.1941874

12.5 Stichprobenbeschreibung PISA 2022

Julia Mang, Sabrina Wagner, Jens Gomolka & Sabine Meinck

Die PISA-Studie zielt darauf ab, Aussagen über Bildungssysteme für fünfzehnjährige Schüler*innen in Deutschland im internationalen Vergleich zu treffen. Die Ziehung der dafür erforderlichen Stichprobe erfolgt anhand exakter statistischer Regeln und erlaubt durch die Verwendung von Stichproben- und sog. Replikatgewichten eine Auswertung der Ergebnisse, die Verallgemeinerungen über alle fünfzehnjährigen Schüler*innen in Deutschland ermöglicht (Kish, 1995; Levy, 2008; Rutkowski et al., 2013). Dieser Ansatz entspricht dem Vorgehen vorheriger Erhebungsrunden. Das vorliegende Kapitel erläutert im Detail die Prozeduren der Stichprobenziehung, Gewichtung und Vorbereitung der Daten zur Varianzschätzung.

12.5.1 Populationsdefinitionen und Stichprobendesign

Um Rückschlüsse aus der stichprobenbasierten PISA-Erhebung auf die Grundgesamtheit der fünfzehnjährigen Schüler*innen aller Teilnehmerstaaten zu ermöglichen sowie die internationale Vergleichbarkeit zu sichern, sind Verfahren der Stichprobenziehung anzuwenden, die unverzerrte und präzise Populationsschätzer ermöglichen (Meinck, 2020). In PISA werden in allen Teilnehmerstaaten zwei- oder mehrstufige Zufallsverfahren für die Ziehung der Stichprobe eingesetzt.[3] In der Regel werden in einem ersten Schritt Schulen gezogen und in einem zweiten Schritt Schüler*innen in den teilnehmenden Schulen systematisch randomisiert ausgewählt. Dieses Verfahren wird auch in Deutschland implementiert. Ein valides Stichproben-Regelwerk gewährleistet und sichert diese Standards in allen teilnehmenden Staaten (OECD, 2020).

In Deutschland wird das Studiendesign für PISA 2022 noch durch eine zusätzliche Zielpopulation erweitert. Mit Hilfe dieser sollen auch Aussagen für *Schüler*innen der 9. Klassenstufe* ermöglicht werden. Da sich beide Zielpopulationen, also Fünfzehnjährige und Neuntklässler*innen, zumindest teilweise überlappen, wird die Stichprobenziehung parallel für beide Populationen durchgeführt und soll im Folgenden detailliert beschrieben werden. Weiterhin wird auch die Stichprobenziehung für Lehrkräfte vorgestellt. Neben den Befragungen der Schüler*innen sowie der Lehrkräfte wurden auch die Eltern der Jugendlichen und die Schulleitungen der teilnehmenden Schulen befragt. Die Teilnehmer*innen dieser Befragungen werden direkt über die Schul- und Schülerziehung bestimmt, weshalb in diesem Bericht nicht weiter auf diese Zielgruppen eingegangen wird. Auch erweitert die Zusatzstudie *Classroom experience, characteristics, and outcome* (Ceco) das Studiendesign für PISA 2022. Hierfür wurden in ausgewählten PISA-Schulen

3 Eine genaue Beschreibung der in den bisherigen PISA-Erhebungsrunden verwendeten Methodologie kann den sogenannten Technical Reports entnommen werden. Diese finden sich auf der Website der OECD – unter: *https://www.oecd.org/pisa/publications/*.

zusätzlich Unterrichtsstunden in Mathematik und in naturwissenschaftlichen Fächern beobachtet sowie Fragen zum Unterricht an diese Lehrkräfte und Schüler*innen gestellt. Auch werden Prüfungsaufgaben dieser Lehrkräfte empirisch im Kontext ausgewertet. Da diese Zusatzstudie nicht Teil der regulären PISA-Erhebung ist, wird sie in weiteren Publikationen ausführlich berichtet.

12.5.1.1 Stichprobe der Fünfzehnjährigen

Wie auch in den vergangenen PISA-Erhebungen besteht die international vorgegebene Zielpopulation aus allen fünfzehnjährigen Schüler*innen, die sich in der siebten oder einer höheren Klassenstufe befinden. Die genaue Definition der Altersgruppe findet in Abstimmung mit dem internationalen PISA-Konsortium statt und kann sich zwischen den Staaten aufgrund verschiedener Erhebungszeiträume leicht unterscheiden. Für Deutschland gilt die folgende Definition der Zielpopulation für PISA 2022:

> *Teilnahmeberechtigt waren alle Schüler*innen, die zwischen dem 1. Januar 2006 und dem 31. Dezember 2006 (einschließlich) geboren sind und die mindestens die 7. Klassenstufe oder eine höhere Klassenstufe besuchen.*

Um vertiefende Analysen durchführen zu können, werden *Schüler*innen der 9. Klassen an allgemeinen Schulen sowie Förderschulen*[4] als weitere Zielpopulation für Deutschland definiert. Schüler*innen in 9. Klassen können aus verschiedenen Altersgruppen stammen, also sowohl Fünfzehnjährige als auch Nichtfünfzehnjährige inkludieren. Somit können Fünfzehnjährige, die eine 9. Klasse besuchen, zumindest theoretisch Teil beider Stichproben sein: die der PISA-Basisstichprobe sowie die der Zusatz-Stichprobe für 9. Klassen. Das unten beschriebene Ziehungsdesign ermöglicht vollumfängliche Repräsentativität für beide Zielpopulationen.

Der vorliegende Bericht enthält vorwiegend Aussagen zur PISA-Grundgesamtheit, basierend auf der Stichprobe der Fünfzehnjährigen. Die Daten der teilnehmenden Neuntklässler*innen werden zu einem späteren Zeitpunkt separat analysiert und berichtet.

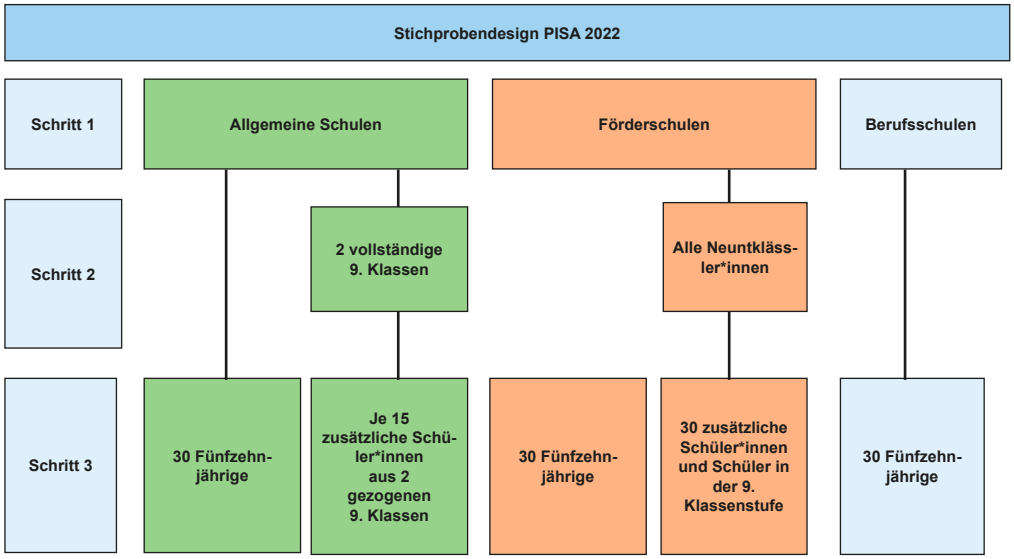
Das Ziehungsverfahren für beide Stichproben setzte sich aus mehreren, teils parallellaufenden Schritten zusammen: Zunächst wurden die Schulen gezogen. Die Liste teilnahmeberechtigter Schulen für beide Schülerpopulationen ist äquivalent – alle Schulen, die aufgrund ihres Typs zumindest theoretisch Fünfzehnjährige bzw. Neuntklässler*innen beschulen können, sind Teil der Ziehungsliste. Diese Schulen bilden somit die Liste der primären Stichprobeneinheiten für beide Zielpopulationen auf Schülerebene. Anschließend erfolgte in einem zweiten Schritt eine Zufalls-Ziehung zweier vollständiger 9. Klassen an allgemeinen Schulen aus allen 9. Klassen der jeweiligen Schule. Falls weniger als drei 9. Klassen vorhanden sind, sind diese Klassen für die Studie gesetzt. An

4 Berufsschulen sind nicht Teil der definierten Schulpopulation für diese Zusatzerhebung.

Förderschulen wurden alle Neuntklässler*innen zu einer einzelnen virtuellen 9. Klasse zusammengruppiert und alle Neuntklässler*innen zur Teilnahme eingeladen.[5] Innerhalb jeder teilnehmenden Schule wurde dann eine Liste aller Fünfzehnjährigen aus den Klassenstufen 7 und höher erstellt und 30 dieser Schüler*innen gezogen. Diese Stichprobe bildet die Teilnehmer*innenschaft für die PISA-Basisstichprobe. Anschließend an diesen Ziehungsschritt wurde ein dritter Ziehungsschritt innerhalb der gezogenen neunten Klassen durchgeführt. In jeder ausgewählten neunten Klasse wurden zunächst die bereits für die PISA-Basisstichprobe ausgewählten Schüler*innen aus der Ziehungsliste entfernt. Danach wurden 15 Schüler*innen per Zufall gezogen. Waren weniger als 16 Schüler*innen gelistet, gelangen automatisch alle Schüler*innen der Liste in die Stichprobe. Alle teilnehmenden Fünfzehnjährigen in der so gezogenen Stichprobe sind Teil des PISA-Datensatzes und werden in den entsprechenden Analysen berücksichtigt. Die verschiedenen Ziehungswahrscheinlichkeiten werden durch entsprechende Designgewichte reflektiert. Das Gleiche gilt für die teilnehmenden Neuntklässler*innen, dem entsprechenden Datensatz der Zusatzstichprobe für neunte Klassen und den entsprechenden Designgewichten. Jede*r fünfzehnjährige Jugendliche, der sich in einer der ausgewählten Klassen der Jahrgangsstufe 9 befindet, hatte zwei Chancen, für die Teilnahme ausgewählt zu werden. Dies wird bei der Gewichtung der Stichprobe der 9. Klassen angemessen berücksichtigt. Die Daten dieser Schüler*innen gehen in die Analysen beider Stichproben, also sowohl in die der PISA-Population als auch in die der Zusatzstichprobe, ein.

Eine Übersicht über das Stichprobendesign bietet die folgende Abbildung 12.5.1.

Abbildung 12.5.1: Stichprobendesign PISA 2022 – Schüler*innen



5 In Förderschulen sind generell weniger Neuntklässler*innen anzutreffen, zudem ist die übliche Klassenstruktur oft nicht gegeben.

12.5.1.2 Lehrkräftestichprobe

Deutschland nimmt wie auch in den vergangenen Erhebungen seit PISA 2015 an der zusätzlichen Lehrkräftebefragung teil.[6] Für PISA 2022 wird erstmals eine Befragung von Lehrkräften in den ausgewählten PISA-Schulen vorgenommen, mit dem Ziel, repräsentative Aussagen über die Lehrkräfte dieser Schulen zu realisieren.

Von besonderem Interesse sind in diesem Zusammenhang Lehrkräfte, die Klassenstufen mit einem hohen Anteil an zur PISA-Grundgesamtheit (Fünfzehnjährige) gehörenden Schüler*innen unterrichten. In Deutschland befinden sich fünfzehnjährige Schüler*innen insbesondere in den Klassenstufen 9 und 10 (Statistisches Bundesamt). Die Definition der Zielpopulation der Lehrkräfte wird daher wie folgt festgesetzt:

> *Die Lehrkräftebefragung ist an alle Lehrkräfte (inkl. Referendare) gerichtet, die aktuell eine 9. und/oder 10. Jahrgangsstufe unterrichten. Vollzeit- sowie Teilzeitlehrkräfte, angestellte und verbeamtete Lehrkräfte sind dabei gleichermaßen zu berücksichtigen. Auch Lehrkräfte, die ihre Lehrtätigkeit an mehreren verschiedenen Schulen ausüben, sind für die Studienteilnahme vorgesehen.*
>
> *Als Lehrkraft gilt dabei eine Person, deren vorrangige oder hauptsächliche Aktivität in der Schule die Ausbildung von Schüler*innen ist und die den Schüler*innen Unterrichtsstunden erteilt. Lehrkräfte können mit den Schüler*innen im ganzen Klassenverband in Klassenräumen arbeiten, in Kleingruppen oder im Einzelunterricht inner- oder außerhalb der regulären Klassenräume.*

Entsprechend der internationalen Definition der Lehrkräftepopulation werden zwei Gruppen von Lehrkräften unterschieden:

1) Lehrkräfte, die das Fach Mathematik unterrichten (da Mathematik in PISA 2022 die Hauptdomäne darstellt).
2) Lehrkräfte, die sonstige Fächer unterrichten.

Die Realisierung der Lehrkräftebefragung erfolgte ebenfalls im Rahmen der Ziehung einer Stichprobe. Die PISA-Schulstichprobe wird auch für die Lehrkräftebefragung genutzt. Innerhalb der teilnehmenden Schulen wurden dann jeweils zehn Lehrkräfte, die in der 9. und/oder 10. Jahrgangsstufe Mathematik unterrichten, im Zufallsverfahren gezogen sowie 15 Lehrkräfte, die sonstige Fächer in diesen Jahrgangsstufen unterrichten.

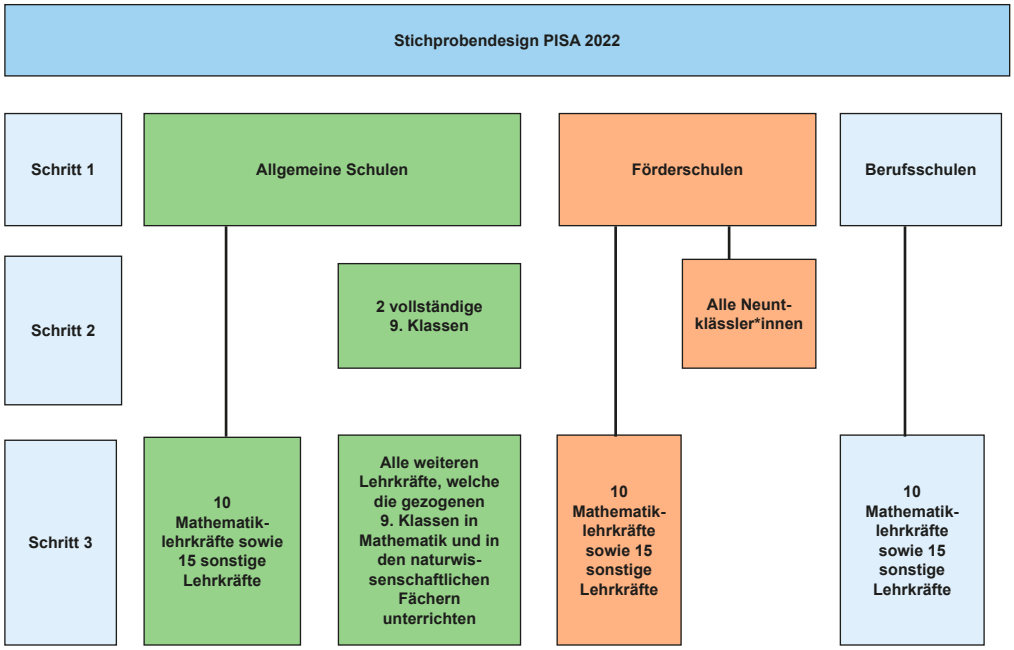
Um auch Analysen zum Klassenkontext der gezogenen 9. Klassen durchführen zu können, wurden Lehrkräfte an allgemeinen Schulen, die die jeweils gezogenen Klassen im laufenden Schuljahr in Mathematik und den naturwissenschaftlichen Fächern unterrichten, nach der Stichprobenziehung zusätzlich in die Lehrkräftestichprobe aufgenommen, wenn sie nicht bereits Teil der Zufallsstichprobe waren.

---

6 Die Lehrkräftebefragung ist erst seit der PISA-Testung im Jahr 2015 Teil der Erhebung. Zwar wurden in Deutschland auch in den vorangegangenen PISA-Zyklen Lehrkräftebefragungen durchgeführt, jedoch nur als rein nationale Zusatzerhebung.

Somit kann die bereits bekannte Darstellung des Stichprobendesigns für PISA 2022 wie folgt angepasst werden:

Abbildung 12.5.2: Stichprobendesign PISA 2022 – Lehrkräfte



12.5.2 Ablauf und Ergebnisse der Stichprobenziehungen

Im folgenden Abschnitt wird erläutert, wie die Ziehungen der Schul-, Schüler*innen- und Lehrkräftestichproben vorbereitet wurden. Außerdem werden die Ergebnisse der Ziehungen vorgestellt.

12.5.2.1 Ziehung der Schulstichprobe

Für die Ziehung der Schulen wird ein sogenannter *Sampling Frame* erstellt. Hierbei handelt es sich um eine umfassende Liste aller Schulen, an welchen potenziell fünfzehnjährige Schüler*innen unterrichtet werden.

Die Informationen zur Erstellung dieses *Sampling Frames* wurden bei den Ansprechpersonen für Schulstatistik bei den statistischen Landesämtern bzw. den jeweiligen Kultusministerien eingeholt. Die Listen umfassen alle Schultypen (allgemeine Schulen, Förderschulen, Berufsschulen) im jeweiligen Land und enthalten die folgenden Angaben:

- die offizielle Schulnummer (die später im *Sampling Frame* pseudonymisiert wurde),
- die Schulart,
- die Anzahl der Schüler*innen in den Geburtsjahrgängen 2005, 2006 und 2007,
- die Anzahl der Schüler*innen in den Klassenstufen 7 bis 10,
- die Anzahl der 7. bis 10. Klassen,
- die Trägerschaft (öffentlich oder privat),
- Informationen über Veränderungen der Schulart, Schulzusammenlegungen und Schulschließungen sowie
- für Förderschulen die Informationen über die Förderschwerpunkte gemäß den Vorgaben der Kultusministerkonferenz (KMK), wobei in Anlehnung an alle vorhergehenden PISA-Erhebungsrunden die Förderschwerpunkte Lernen, Sprache sowie emotionale und soziale Entwicklung berücksichtigt wurden.

Als Datengrundlage für die genannten Schulinformationen dienten die Schulstatistiken der einzelnen Länder für das Schuljahr 2020/2021. Falls diese für eine oder mehrere Informationen nicht verfügbar waren, wird auf die jeweils zuletzt veröffentlichten Daten zurückgegriffen.

PISA wendet zur Zufallsziehung der Schulen das sogenannte PPS-Verfahren (*Probabilities Proportional to Size*; zum Beispiel Skinner, 2014) an. Hierbei wird die Ziehungswahrscheinlichkeit umgekehrt proportional zur Schulgröße festgesetzt. Große Schulen haben somit eine erhöhte Wahrscheinlichkeit, gezogen zu werden. Umgekehrt haben Schüler*innen innerhalb großer Schulen eine kleine Wahrscheinlichkeit, für die Studie ausgewählt zu werden. Diese Methode führt zu geringen Varianzen in den Stichprobengewichten und trägt somit zu niedrigen Standardfehlern bei. Um dieses Ziehungsverfahren anwenden zu können, muss der *Sampling Frame* ein sogenanntes *Measure of Size* (MOS) aufweisen. In PISA stellt die erwartete Anzahl an fünfzehnjährigen Schüler*innen pro Schule ein optimales MOS dar.

In Ermangelung eines exakten Wertes wurde dem *Sampling Frame* ein Schätzer der zu erwartenden Anzahl an Fünfzehnjährigen pro Schule im Jahr 2022 hinzugefügt, um die Durchführung des in PISA präferierten PPS-Ziehungsverfahrens möglich zu machen. Als Schätzer wurde die jeweilige Schüleranzahl des Geburtsjahrgangs 2005 verwendet, welche außerdem mit den Schülerzahlen der Jahrgänge 2004 und 2003 abgeglichen wurde, um Schwankungen in den Geburtenzahlen zu identifizieren.[7]

Um Aussagen zum Stichprobenumfang hinsichtlich des erweiterten Forschungsdesigns zu ermöglichen, wird im *Sampling Frame* außerdem die Anzahl zu erwartender Neuntklässler*innen gelistet. Auch hierfür muss ein Schätzer verwendet werden, wobei es sich um die Anzahl der Neuntklässler*innen an der jeweiligen Schule im Schuljahr

7 Es wird nicht der Geburtsjahrgang 2006 (PISA-Zielpopulation) verwendet, da so die Anzahl an Schüler*innen, die sich 2022 an allgemeinen Schulen befinden, überschätzt werden könnte: Da in einigen Bundesländern viele Fünfzehnjährige die Hauptschule verlassen, um berufsbildende Schulen zu besuchen, dürfen nicht die Vierzehnjährigen als Schätzer für die Fünfzehnjährigen des nächsten Schuljahres verwendet werden.

2020/2021 handelte. Um mögliche Inkonsistenzen ausgleichen zu können, wurden auch die Schülerzahlen der 8. und 10. Klassenstufe überprüft.

Außerdem wurden alle verwendeten Daten der statistischen Landesämter mit den veröffentlichten Daten des Statistischen Bundesamtes abgeglichen, um möglichen Fehlern oder Ungenauigkeiten entgegenzuwirken und somit die Qualität der Stichprobenziehung zu optimieren. Auftretende Auffälligkeiten wurden in Rücksprache mit dem jeweiligen statistischen Landesamt geklärt.

Bevor die Stichprobenziehung der Schulen durchgeführt werden kann, müssen die Schulen im *Sampling Frame* nach bestimmten Kriterien, den sogenannten Stratifizierungsvariablen, gruppiert werden.

Eine Gruppe wird als Stratum bezeichnet und enthält einander „ähnliche" Schulen, weil eben alle Schulen eines Stratums das durch die Stratifizierung bestimmte Merkmal miteinander teilen. Wenn mehrere Stratifizierungsvariablen verwendet und verknüpft werden, teilen Schulen eines Stratums entsprechend mehrere gemeinsame Merkmale. Die Stratifizierung kann explizit und implizit vorgenommen werden. Bei der expliziten Stratifizierung werden die Schulen in einzelne Gruppen beziehungsweise Strata aufgeteilt, welche unabhängig voneinander behandelt werden. Aus jedem Stratum wird dann eine separate Zufallsauswahl getroffen. Bei der impliziten Stratifizierung werden die Schulen innerhalb eines expliziten Stratums im *Sampling Frame* noch einmal sortiert, um eine näherungsweise proportionale Verteilung der gezogenen Schulen über die impliziten Strata innerhalb des expliziten Stratums zu gewährleisten.

In Deutschland kommen für PISA 2022 jeweils zwei explizite und implizite Stratifizierungsvariablen zum Einsatz:

Zunächst werden alle Schulen, welche potenziell fünfzehnjährige Schüler*innen unterrichten, in drei Strata aufgeteilt: allgemeine Schulen, Förderschulen und Berufsschulen (1. explizite Stratifizierung). Anschließend werden die allgemeinen Schulen erneut in Gruppen aufgeteilt, und zwar den 16 Bundesländern entsprechend (2. explizite Stratifizierung). Damit ergeben sich insgesamt 18 explizite Strata. 16 davon repräsentieren allgemeine Schulen in den einzelnen Bundesländern, ein Stratum enthält alle Förderschulen und eines enthält alle Berufsschulen. Die Förder- und Berufsschulen werden gesondert behandelt, da die relative Häufigkeit dieser Schulformen innerhalb der Bundesländer sehr unterschiedlich ist. Auch die im Verhältnis zu den anderen Schularten sehr geringe Anzahl an Schüler*innen an Förder- und Berufsschulen erfordert die Berücksichtigung der beiden Schulformen als explizite Strata.

Weiterhin werden die allgemeinen Schulen in die Schulformen Hauptschule, Integrierte Gesamtschule, Realschule, Schule mit mehreren Bildungsgängen, Gymnasium und Schule-nicht-deutscher-Herkunftssprache[8] eingeteilt (1. implizite Stratifizierung). Diese Maßnahme stellt sicher, dass sich die gezogenen allgemeinen Schulen innerhalb der ein-

8 Die Schulform ND wird ausschließlich im Bundesland Hessen gesondert ausgewiesen. Dies umfasst Schüler*innen aus dem Ausland, die grundlegende Kenntnisse der deutschen Sprache erwerben müssen und nicht in Regelklassen unterrichtet werden. In den übrigen Bundesländern sind diese Schüler*innen den Schularten bzw. Bildungsbereichen zugeordnet.

zelnen Bundesländer näherungsweise proportional zur Gesamtzahl der Schüler*innen auf die verschiedenen Schulformen verteilen. Um auch aus allen Bundesländern eine annähernd proportionale Menge zur tatsächlichen Anzahl an Schüler*innen an Berufs- und Förderschulen in der Stichprobe zu haben, werden diese beiden Strata außerdem jeweils nach Bundesländern geschichtet (2. implizite Stratifizierung).

Anhand dieser Vorgehensweise wird sichergestellt, dass bei der anschließenden Ziehung der Schulstichprobe vom internationalen Konsortium eine Stichprobe gezogen wird, welche Schulen aus allen 18 Strata enthält und sich hinsichtlich der Anzahl an Schüler*innen pro Bundesland und Schulform annähernd proportional zur Grundgesamtheit (Geburtenjahrgang 2006) verhält.

Insgesamt wurden 272 Schulen für PISA 2022 gezogen. Tabelle 12.5.1 zeigt die Zuordnung gezogener Schulen zu Bundesländern und Schularten:

Tabelle 12.5.1: Zuordnung gezogene Schulen zu Schularten und Bundesländer

Bundesland	HS	RS	MBG	G	IG	ND	F	B	∑
Baden-Württemberg	3	12	0	12	6	0	2	5	***40***
Bayern	11	13	0	12	1	0	2	4	***43***
Berlin	0	0	0	4	6	0	1	0	***11***
Brandenburg	0	0	3	3	1	0	0	1	***8***
Bremen	0	0	0	0	2	0	0	0	***2***
Hamburg	0	0	0	2	3	0	0	0	***5***
Hessen	2	4	0	9	4	2	1	1	***23***
Mecklenburg-Vorpommern	0	0	2	2	0	0	1	0	***5***
Niedersachsen	2	3	6	9	4	0	1	1	***26***
Nordrhein-Westfalen	3	12	3	19	17	0	3	2	***59***
Rheinland-Pfalz	0	1	4	4	2	0	1	0	***12***
Saarland	0	0	0	1	2	0	0	1	***4***
Sachsen	0	0	7	4	0	0	1	1	***13***
Sachsen-Anhalt	0	0	3	2	1	0	0	0	***6***
Schleswig-Holstein	0	0	0	2	5	0	1	1	***9***
Thüringen	0	0	3	2	1	0	0	0	***6***
Gesamt	21	45	31	87	55	2	***14***	***17***	272

Anmerkung: Hervorgehobene Zahlen entsprechen den expliziten Strata; HS (Hauptschule), RS (Realschule), MBG (Schule mit mehreren Bildungsgängen), G (Gymnasium), IG (integrierte Gesamtschule), ND (Schule nicht deutscher Herkunftssprache), F (Förderschule), B (Berufsschule).

Wie in Tabelle 12.5.1 dargestellt, wird nicht aus jedem Stratum auch jede Schulform gezogen. Alle Schulen haben jedoch eine von Null verschiedene Ziehungswahrscheinlichkeit.[9]

12.5.2.2 Ziehung der Schüler*innenstichproben

Nach der Bestimmung der Schulen, an denen die Datenerhebung durchgeführt werden soll, können dem oben vorgestellten Stichprobendesign entsprechend die Schüler*innen-, Klassen- und Lehrkräftestichproben gezogen werden.

Um eine Schüler*innenstichprobe zu erhalten, die die Anforderungen des Studiendesigns erfüllt, müssen verschiedene demografische Daten herangezogen werden. Von den für PISA 2022 gezogenen Schulen wurden alle teilnahmeberechtigten Schüler*innen, also sowohl alle Fünfzehnjährigen als auch alle Neuntklässler*innen, gelistet. Diese Schüler*innenlisten enthalten unter anderem die Merkmale Geburtsjahr, Geschlecht, Klassenstufe und Klassenbezeichnung. Personenbezogene Daten wie beispielsweise die Namen der Schüler*innen verbleiben dabei in den Schulen, um volle Pseudonymität sicherzustellen.

Beruhend auf diesen Angaben können außerdem Auflistungen aller 9. Klassen pro Schule bereitgestellt werden.

Für den Umgang mit diesen Listen wurde das von der *International Association for the Evaluation of Educational Achievement Hamburg* (IEA Hamburg) entwickelte Online-System *IEA OnlineSurveyExpert* (IEA OSE) verwendet. Über dieses System können Informationen zwischen den Schulen und dem Datenerhebungsinstitut, der IEA Hamburg, ausgetauscht und gleichzeitig alle datenschutzrechtlichen Belange durch eine verschlüsselte Übertragung sowie die Pseudonymisierung persönlicher Daten berücksichtigt werden.

Den Schulen lag dementsprechend eine vollständige Liste mit allen relevanten schülerbezogenen Daten, inklusive Namensangaben, vor. Der IEA Hamburg hingegen lag die gleiche Liste nur in pseudonymisierter Form vor, welche anstelle von Namen Ordnungsnummern enthält[10].

Für die Stichprobenziehung von Klassen und Schüler*innen kommt die vom internationalen Konsortium bereitgestellte Software *Maple* zur Anwendung. Zunächst werden die Klassenlisten in die Software eingelesen und die Klassenziehung durchgeführt. Für Förderschulen werden wie bereits erwähnt alle Neuntklässler*innen in einer virtuellen Klasse zusammengefasst. Nachdem so je Schule zwei 9. Klassen gezogen wurden,

9 Ausführliche Beispiele zur Schulstichprobenziehung sowie den dazugehörigen Ziehungsalgorithmen können den PISA Technical Reports unter: *https://www.oecd.org/pisa/publications/* entnommen werden.

10 Die Ordnungsnummern sind auch in den Listen, die den Schulen vorlagen, enthalten und dem entsprechenden Namen zugeordnet. Somit erfolgt jegliche Kommunikation zwischen den Schulen und der IEA Hamburg ausschließlich über diese Ordnungsnummern.

wurden auch die Schüler*innenlisten in Maple eingelesen und die Stichprobe der 30 Fünfzehnjährigen pro Schule gezogen. Anschließend wurden an allgemeinen- und Förderschulen zusätzlich jeweils 15 Neuntklässler*innen innerhalb der zuvor gezogenen 9. Klassen per Zufallsziehung ermittelt.[11]

So wurden n = 7206 Schüler*innen für die Testgruppe der Fünfzehnjährigen sowie weitere n = 8605 für die Testgruppe der Neuntklässler*innen ermittelt.[12]

12.5.2.3 Ziehung der Lehrkräftestichprobe

Auch die Ziehung der Lehrkräfte erfolgt innerhalb der für PISA 2022 ausgewählten Schulen. Ebenso wie für die Ermittlung der Stichproben auf Schüler*innenebene wird für die Ziehung der Lehrkräfte von den jeweiligen Schulen eine Liste erstellt, in diesem Fall die Lehrer*innenliste, welche der IEA Hamburg ebenfalls in pseudonymisierter Form vorliegt. Diese enthält wie auch die Schüler*innenliste demografische Merkmale sowie Angaben zu den unterrichteten Fächern. Auch diese Liste wurde in das Stichprobenziehungsprogramm Maple importiert, sodass anschließend die Stichprobenziehung gemäß dem bereits vorgestellten Stichprobendesign für Lehrkräfte durchgeführt werden kann.

Durch dieses Verfahren ergab sich eine Gesamtzahl von n = 1499 Mathematik-Lehrkräften. Davon wurden n = 1456 zufällig gezogen und n = 43 weitere Mathematik-Lehrkräfte zusätzlich in die Stichprobe aufgenommen, da sie eine der beiden gezogenen 9. Klassen unterrichten. Bei den sonstigen Lehrkräften ergab sich eine Gesamtzahl von n = 3820, von denen n = 3552 zufällig gezogen wurden sowie n = 268 zusätzlich aufgenommen wurden. Damit liegt die Gesamtzahl der Lehrkräftestichprobe für PISA 2022 bei n = 5319.

Schließlich wurden die Ergebnisse der Klassen-, Schüler*innen- und Lehrerkräftestichprobenziehungen an die Schulen weitergegeben, sodass diese die Durchführung der Testung vorbereiten können.

12.5.3 Teilnahmequoten (realisierte Stichproben)

Bei den bisher beschriebenen Stichproben handelt es sich um sogenannte Bruttostichproben. Sie umfassen alle für die Teilnahme an PISA 2022 ausgewählten Schulen, Schüler*innen sowie Lehrkräfte. Hiervon zu unterscheiden sind die Nettostichproben, welche lediglich die Schulen und Personengruppen umfassen, die auch tatsächlich an der Studie teilgenommen haben und von denen auswertbare Daten vorliegen.

---

11 Für die Berufsschulen ist die Klassenziehung nicht relevant, da an diesen Schulen nur die Testgruppe der fünfzehnjährigen Schüler*innen gezogen wird.

12 Die gezogenen Neuntklässler*innen, welche auch der Gruppe der Fünfzehnjährigen zuzuordnen sind, werden für Analysen in beiden Gruppen herangezogen.

Die Brutto- und Nettostichproben in PISA 2022 unterscheiden sich wie folgt: An acht von 272 insgesamt gezogenen Schulen gab es keine fünfzehnjährigen Schüler*innen, sodass keine Testung stattfinden konnte. Bei diesen acht Schulen handelte es sich um vier Berufsschulen, eine Förderschule, eine Grundschule sowie drei Schulen, die sich im Aufbau befinden.[13] Gemäß den internationalen PISA-Standards dürfen solche Schulen nicht durch andere ersetzt werden, da sie für PISA nicht teilnahmeberechtigt sind.[14] Sie haben keinen Einfluss auf die Teilnahmeraten. Zwei weitere Schulen wurden im Verlauf der Studie ausgeschlossen und ebenfalls nicht substituiert. Weiterhin verweigerte eine Schule die Teilnahme und konnte durch keine ihrer beiden Ersatzschulen ersetzt werden. Somit ist auch diese Schule in der Nettostichprobe nicht enthalten. 11 Schulen wurden hingegen jeweils durch ihre erste Ersatzschule und fünf Schulen durch ihre zweite Ersatzschule ersetzt. Werden also von der Bruttostichprobe $n = 272$ die acht nicht teilnahmeberechtigten Schulen abgezogen, ergibt sich eine korrigierte Bruttoschulstichprobe von $n = 264$ Schulen. Hiervon haben 257 Schulen teilgenommen. Dies entspricht einer ungewichteten Teilnahmerate von 91.3 Prozent basierend auf den original gezogenen Schulen und 97.3 Prozent nach Berücksichtigung der Ersatzschulen.

Die gewichtete Teilnahmerate auf Schulebene beträgt 92.9 Prozent ohne und 98.2 Prozent unter Berücksichtigung von Ersatzschulen.

Auch auf der Ebene der Schüler*innen kam es zu Ausfällen. Hierfür gibt es verschiedene Gründe. So kann es vorkommen, dass einzelne Jugendliche aufgrund eines sonderpädagogischen Förderbedarfs offiziell von der Studie ausgeschlossen wurden. Da ein solcher Ausschluss bereits vor der Testung erfolgt, werden diese Ausfälle in der Teilnahmequote nicht berücksichtigt. Weitere Ausfälle sind durch Erkrankungen der Schüler*innen oder kurzfristig erfolgte Schulwechsel zu erklären.[15] Es traten vor allem erhöhte krankheitsbedingte Ausfälle der Schüler*innen während der Corona-Pandemie in allen teilnehmenden Schulen auf, sodass eine Vielzahl von Nachtestungen an den Schulen notwendig war. Mit Hilfe des engagierten Einsatzes aller Beteiligten in der Feldarbeit, explizit der Schulkoordinator*innen, der Testleitungen sowie der Unterstützung politischer Träger konnten alle Testungen erfolgreich durchgeführt werden. Damit ergab sich eine korrigierte Bruttoschülerstichprobe von $n = 6964$ für die Testgruppe der fünfzehnjährigen gezogenen Schüler*innen. Davon nahmen insgesamt $n = 6116$ Schüler*innen teil, woraus sich eine ungewichtete Teilnahmerate von 87.8 Prozent ergibt. Die gewichtete Teilnahmequote auf Schülerebene liegt bei 88.0 Prozent.

---

13 Fälle wie diese sind darauf zurückzuführen, dass die amtlichen Daten zur Schulstatistik zu einem bestimmten Zeitpunkt erhoben werden, die Stichprobenziehung jedoch erst später erfolgt und die deutsche Schullandschaft teilweise recht dynamisch ist.

14 Diese Schulen repräsentieren andere Schulen in Sampling Frame, die ebenfalls fälschlicherweise auf der Schulliste standen.

15 Letzteres kann vorkommen, da zwischen dem Zeitpunkt der Schülerlistung an den Schulen und der Durchführung der Stichprobenziehung mehrere Wochen vergehen können, in denen sich – wenn auch selten – die Zusammensetzung der Schüler*innen in den Schulen verändern kann (zum Beispiel aufgrund von Wegzügen einzelner Schüler*innen).

Diese Teilnahmequoten lassen auf valide und vollumfänglich repräsentative Ergebnisse schließen. Auf die zusätzliche Testgruppe der Neuntklässler*innen wird – wie bereits erwähnt – in diesem Bericht nicht weiter eingegangen.

Da der Lehrkräftefragebogen nicht in allen Bundesländern verpflichtend war, ist hier eine geringere Teilnahmequote als bei den Schüler*innen zu beobachten. Von den insgesamt n = 5319 für PISA 2022 ausgewählten Lehrkräften nahmen n = 3899 an der Befragung teil. Die ungewichtete Teilnahmequote liegt damit bei 73.3 Prozent und somit immer noch auf akzeptablem Niveau. Wird dabei zwischen den Quoten der Mathematik- und sonstigen Fachlehrkräfte differenziert, findet sich eine leicht höhere Teilnahme für die Mathematiklehrkräfte. Die ungewichtete Teilnahmequote dieser Lehrkräfte liegt bei 76.3 Prozent, die der sonstigen Lehrkräfte bei 72.7 Prozent.

12.5.4 Gewichtung und Nichtteilnahmeadjustierung als Reflektion unterschiedlicher Ziehungswahrscheinlichkeiten und Nichtteilnahmemuster

Aufgrund des komplexen Designs der Stichprobenziehung haben nicht alle Schüler*innen die gleiche Wahrscheinlichkeit, gezogen zu werden. Dazu kommt, dass nicht alle gezogenen Schüler*innen auch tatsächlich an der Testung teilnehmen (vgl. Abschnitt 12.5.3). Das Vorliegen gleicher Ziehungswahrscheinlichkeiten für jede Untersuchungseinheit (Schüler*innen) und die Teilnahme bei erfolgter Ziehung sind aber notwendige Voraussetzungen für die Verallgemeinerbarkeit von Stichprobenergebnissen auf die Zielpopulation (Bortz & Döring, 2016). Um die ungleichen Ziehungswahrscheinlichkeiten sowie auch unterschiedliche Teilnahmeraten auszugleichen, werden Schulbasis- beziehungsweise Schüler*innenbasisgewichte sowie verschiedene Korrekturfaktoren verwendet. Die Basisgewichte werden separat für jeden Ziehungsschritt umgekehrt proportional zur Ziehungswahrscheinlichkeit errechnet. Haben also beispielsweise Schulen einer Schulform aufgrund des Stichprobendesigns eine geringere Wahrscheinlichkeit, gezogen zu werden, ergibt sich entsprechend für diese Schulen ein höheres Schulbasisgewicht. Die Schüler*innenbasisgewichte innerhalb der gezogenen Schulen werden ebenfalls umgekehrt proportional zur Wahrscheinlichkeit der Schüler*innenziehung errechnet. Hier haben Schüler*innen einer großen Schule eine geringere Wahrscheinlichkeit, gezogen zu werden als Schüler*innen einer kleinen Schule. Daher erhalten die Jugendlichen, die eine große Schule besuchen, ein höheres Schüler*innenbasisgewicht.

Weiterhin gehen fünf Korrekturfaktoren in die Gewichtung ein. Zunächst muss der Ausfall von Schulen berücksichtigt werden. Sollte es zu einem Schulausfall kommen, werden andere Schulen, die derselben expliziten Schicht angehören und somit der ausgefallenen Schule möglichst ähnlich sind, höher gewichtet, um den Ausfall zu kompensieren (1. Korrekturfaktor). Auch auf der Ebene der Schüler*innen wird das Basisgewicht um deren Nichtteilnahme korrigiert (2. Korrekturfaktor). Damit wird vermieden, dass es zu einer Über- oder Unterrepräsentation von Jugendlichen bestimmter Subpopu-

lationen kommt. Zwei weitere Korrekturfaktoren gleichen Differenzen der Schulbasis- beziehungsweise Schüler*innenbasisgewichte zwischen der Stichprobenziehung und der tatsächlichen Größe der Ziehung zum Zeitpunkt der Erhebung aus. Ein weiterer Korrekturfaktor betrifft Staaten, in denen nur die fünfzehnjährigen Schüler*innen befragt werden, welche sich in der Klassenstufe mit der am höchsten zu erwartenden Anzahl an Fünfzehnjährigen befinden. Detaillierte Angaben zur Berechnung der Gewichte können dem Technical Report der OECD zu PISA 2022 entnommen werden (OECD, in Vorbereitung).

Aus den Schulbasis- und Schüler*innenbasisgewichten sowie den fünf Korrekturfaktoren wird durch Multiplikation das Schüler*innengesamtgewicht berechnet. Dieses Schüler*innengesamtgewicht wird für sämtliche Analysen der PISA-Daten auf Ebene der Schüler*innen verwendet, sodass die Ergebnisse für die gezogene Stichprobe auf die gesamte Zielpopulation (fünfzehnjährige Schüler*innen) verallgemeinert werden können.

12.5.5 Präzision der PISA-Daten

Durch die zweistufige Stichprobenziehung ergibt sich bei allen Analysen eine (statistische) Abhängigkeit der Schüler*innen innerhalb der Schulen. Das bedeutet, dass die Merkmale der Jugendlichen (zum Beispiel deren Kompetenzen) innerhalb einer Schule ähnlicher ausfallen als zwischen unterschiedlichen Schulen. Des Weiteren unterscheiden sich diese Abhängigkeiten auch zwischen den einzelnen Teilnehmerstaaten (zum Beispiel aufgrund unterschiedlicher Bildungssysteme), was wiederum unterschiedliche Auswirkungen auf die Schätzgenauigkeit haben kann (vgl. OECD, 2022). Der Effekt des Stichprobendesigns auf den Fehler der Populationsschätzungen (dem sogenannten Standardfehler) wird bei PISA und anderen Schulleistungsstudien als *Designeffekt* bezeichnet.

Da die Größe des Designeffekts bei PISA im Vorfeld nicht eindeutig zu quantifizieren ist und darüber hinaus auch zwischen den einzelnen Teilnehmerstaaten unterschiedlich ausfällt, wird der Standardfehler in PISA seit der ersten Erhebung im Jahr 2000 mit sogenannten Replikationsmethoden auf Basis der erhobenen Daten berechnet. Bei PISA kommt die *Balanced Repeated Replication* (BRR, zum Beispiel Wolter, 2003) zur Anwendung mit einer Erweiterung nach Fay (1989; vgl. dazu auch Judkins, 1990). Die statistische Herleitung der Schätzung des Standardfehlers sowie eine detailliertere Darstellung in Bezug auf die PISA-Studie ist dem Technical Report zu PISA 2018 (OECD, 2022) zu entnehmen und wird für PISA 2022 im Technical Report der OECD erwartet (OECD, in Vorbereitung).

Durch den Einsatz von Replikationsmethoden werden alle Merkmale des Stichprobendesigns bei der Schätzung der Standardfehler berücksichtigt. Damit wird einer möglichen Unterschätzung der Standardfehler vorgebeugt (vgl. Mang et al., 2019; OECD, 2022). Bei der Datenauswertung wird die Berechnung der Varianz durch die Verwendung sogenannter Replikationsgewichte im Datensatz praktisch realisiert (OECD, 2022).

Literatur

Bortz, J., & Döring, N. (2016). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-642-41089-5

Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*(6), 223–239.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. Proceedings of the Survey Research Methods Section of the American Statistical Association, 212–217.

Kish, L. (1995). *Survey sampling*. Wiley & Sons.

Levy, P. S. (2008). *Sampling of populations: Methods and applications* (4. Aufl.). Wiley. https://doi.org/10.1002/9780470374597

Mang, J., Wagner, S., Gomolka, J., Schäfer, A., Meinck, S., & Reiss, K. (2019). *Technische Hintergrundinformationen PISA 2018*. https://mediatum.ub.tum.de/1518258

Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Hrsg.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (S. 113–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_7

OECD. (2020). *PISA 2022 technical standards.* OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2022-Technical-Standards.pdf

OECD. (2022). *PISA 2018 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (in Vorbereitung). *PISA 2022 Technical Report*. OECD Publishing.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of international large-scale Assessment*. CRC Press. https://doi.org/10.1201/b16061

Skinner, C. J. (2014). Probability Proportional to Size (PPS) sampling. *Wiley StatsRef: Statistical Reference Online*, 1–5. https://doi.org/10.1002/9781118445112.stat03346.pub2

Statistisches Bundesamt. *Fachserie 11: Reihe 1: Allgemeinbildenden Schulen*. https://www.destatis.de/DE/Service/Bibliothek/_publikationen-fachserienliste-11.html#558596

Wolter, K. M. (2003). *Introduction to variance estimation.* Springer.

12.6 Testorganisation und Durchführung im Feld

Carola Bretsch, Nina Hugk, Julia Mang & Jörg-Henrik Heine

In diesem Abschnitt werden die organisatorischen und praktischen Aspekte der Durchführung der PISA-Studie in Deutschland dokumentiert. Beschrieben werden die Vorbereitungen im nationalen Projektzentrum an der TU München, der IEA-Hamburg und an den Schulen in ganz Deutschland. Die Beschreibung eines exemplarischen Ablaufs eines PISA-Testtages an den Schulen dokumentiert die praktische Durchführung der Datenerhebung für PISA 2022 in Deutschland. Abschließend wird das Datenmanagement skizziert.

12.6.1 Vorbereitung der Testsitzung

An der Vorbereitung der Testsitzungen sind viele unterschiedliche Akteure beteiligt. Während das PISA-Team an der Technischen Universität München (TUM) die wissenschaftliche Organisation, Konzeption und Begleitung der Studie übernimmt, führt das beauftragte Erhebungsinstitut die *International Association for the Evaluation of Educational Achievement* am Standort in Hamburg (IEA Hamburg) die Datenerhebung an den Schulen durch. Die verantwortliche Feldabteilung der IEA Hamburg koordiniert und organisiert den Kontakt zu den ausgewählten PISA-Schulen und betreut diese über den gesamten Projektverlauf hinweg. Jede an der PISA-Haupterhebung 2022 beteiligte Schule benennt eine Schulkoordinatorin bzw. einen Schulkoordinator, die bzw. der für alle wichtigen Abstimmungsprozesse im Vorfeld der Befragung innerhalb der Schule verantwortlich ist. Dies ist in einigen Fällen die Schulleitung, in der Regel aber ein mit der Schulorganisation gut vertrautes Mitglied des Kollegiums. Die Schulkoordinator*innen fungieren als Ansprechpersonen für die IEA Hamburg sowie für die Testleiter*innen. Zu ihren wichtigsten Aufgaben im Vorbereitungsprozess zählen u. a. die Listung der Fünfzehnjährigen und der Neuntklässler*innen sowie der Lehrkräfte, die in den Jahrgangsstufen 9 und 10 unterrichten, die Festlegung der Testtage im Schulkalender, die Feststellung der verfügbaren Computer an der Schule, die Verteilung der Informationsschreiben an die Schüler*innen und deren Erziehungsberechtigte sowie die Verteilung der Log-in-Briefe an die ausgewählten Lehrkräfte und die Schulleitung. Die Testsitzungen selbst werden von ausgebildeten Testleiter*innen durchgeführt. Diese werden in einem mehrstufigen Schulungsprozess intensiv für die reibungslose Durchführung des Tests von der IEA geschult und ausgebildet. Sie stehen auch im engen und regelmäßigen Austausch mit der Feldabteilung und der Schulkoordinator*in rund um den Testtag.

12.6.2 Exemplarischer Ablauf eines PISA-Testtages

Die PISA-Testsitzungen werden nach Möglichkeit an einem Vormittag in der Schule durchgeführt. Am Testtag treffen sich die Testleiter*innen etwa eine Stunde vor der vereinbarten Startzeit für die Testsitzungen mit der Schulkoordinatorin bzw. dem Schulkoordinator, um letzte Absprachen zu treffen und um die Testräume vorzubereiten.

Im Regelfall gibt es vier Testgruppen pro Schule mit jeweils 15 Schüler*innen. Zwei Testgruppen setzen sich aus Schüler*innen der PISA-Basisstichprobe der Fünfzehnjährigen zusammen, in den anderen beiden Testgruppen befinden sich Schüler*innen der Zusatzstichprobe der 9. Klassen. Je ein*e Testleiter*in führt in separaten Räumen durch die Testsitzung. Um den Vorgaben im Rahmen der gesetzlichen Aufsichtspflicht zu genügen, stellt jede Schule für jede Testgruppe eine Aufsichtsperson aus dem Kollegium, die über den gesamten Verlauf der Testsitzung im Raum anwesend ist.

Am Testtag übergibt die Schulkoordinatorin bzw. der Schulkoordinator den Testleiter*innen für jede Testgruppe eine über die Online-Anwendung IEA OSE (vgl. 11.5.2.2) ausgedruckte Schüler*innenauswahlliste mit den Klarnamen der Schülerinnen und Schüler und die in IEA OSE an jede*n Schüler*in automatisch vergebene fortlaufende Ordnungsnummer. Außerdem erhält die Testleitung einen von der Schulkoordination vorbereiteten Block mit Klebezetteln, auf denen nacheinander die Namen und Ordnungsnummern der in der Testgruppe befindlichen Schüler*innen entsprechend ihrer Reihenfolge auf der Schüler*innenauswahlliste notiert werden. Die ausgedruckte Schüler*innenauswahlliste und die Klebezettel dienen ausschließlich der korrekten Zuordnung der Testinstrumente am Testtag und verbleiben nach Abschluss der Vorbereitungen an der Schule. Auf Grundlage der Ordnungsnummern wurde in IEA OSE für jede*n Schüler*in eine Schüler*innen-Identifikationsnummer (Schüler-ID) erzeugt, die eine Zusammenführung der zur selben Person gehörenden Angaben ermöglicht und das Prinzip der Pseudonymisierung der Daten sicherstellt.

In einigen Bundesländern müssen Erziehungsberechtigte für die Teilnahme ihres Kindes an der Erhebung (nur an Schulen in freier Trägerschaft) beziehungsweise für die Teilnahme am Fragebogen schriftlich einwilligen (s. Tabelle 12.2.2). In diesen Fällen erhalten die Testleiter*innen zusätzlich am Testmorgen von der Schulkoordination die ausgefüllten und unterschriebenen Einwilligungsblätter. Nach Überprüfung der Einwilligungsblätter und der Feststellung, welche Schüler*innen an den Test- und/oder Fragebogenteilen teilnehmen dürfen, gibt die Testleitung die Einwilligungsblätter zurück an die Schulkoordination.

Die Testleiter*innen nutzen die Vorbereitungszeit in den Testräumen, um die Arbeitsplätze für die Schüler*innen einzurichten. Ihnen obliegt die korrekte Zuweisung der Testinstrumente zu den Schüler*innen. Für die Dauer der Vorbereitungen befinden sich noch keine Schüler*innen in den Testräumen. Die Testleiter*innen bringen sämtliche Instrumente und das im Rahmen der Testsitzung benötigte Material mit an die Schule. Hierzu zählen auch die verschlüsselten USB-Sticks, auf denen sich die Testanwendung befindet, sowie gegebenenfalls einzusetzende Laptops. Diese kommen zum

Einsatz, wenn sich im Rahmen der vorab durchgeführten Überprüfung der schuleigenen Computersysteme (Systemdiagnose) gezeigt hat, dass die schuleigenen Computer die technischen Voraussetzungen nicht erfüllen oder Schulcomputer nicht in ausreichender Anzahl zur Verfügung stehen.

In der Vorbereitungsphase werden von den Testleiter*innen zunächst die mitgebrachten Laptops beziehungsweise die schuleigenen Geräte aufgebaut und gestartet. Jedes Gerät wird einem Jugendlichen für den gesamten Verlauf der Testsitzung fest zugeteilt. Im Rahmen der Erhebungsvorbereitung wird in der IEA Hamburg für jede Testgruppe eine Schüler*innenanwesenheitsliste in Papierform erzeugt, die ebenfalls die Ordnungsnummern und Schüler-IDs enthält und die der richtigen Administration der Test- und Befragungsinstrumente sowie der Dokumentation des Teilnahmestatus der Schüler*innen am Testtag dient. Die Zuweisung der Geräte geschieht auf Basis der Login-Blätter, die von der IEA Hamburg im Zuge der Ziehung der Schüler*innenstichprobe erzeugt wurden und die für die teilnehmenden Schüler*innen neben der Schüler-ID die individualisierten Anmeldeinformationen für die Testanwendung enthalten. Mithilfe der ausgedruckten Schüler*innenauswahlliste weisen die Testleiter*innen den einzelnen Schüler*innen die entsprechenden Log-in-Blätter zu. Nachdem die passenden Klebezettel mit dem Klarnamen und der Ordnungsnummer des*der Schüler*in auf das entsprechende Log-in-Blatt geklebt wurden, verteilen die Testleiter*innen die Log-in-Blätter in der Reihenfolge der Schüler-IDs auf die Plätze im Erhebungsraum und melden anschließend jede*n Schüler*in mit den auf dem Log-in-Blatt eingedruckten Anmeldeinformationen am Testprogramm an.

Im letzten Vorbereitungsschritt verteilen die Testleiter*innen noch die Befragungsinstrumente für die Erziehungsberechtigten auf die Arbeitsplätze der Schüler*innen. Hierbei handelt es sich um einen Umschlag, in dem sich ein Fragebogen für die Erziehungsberechtigten, ein Informationsschreiben und ein Rückumschlag befinden. Die Zuordnung erfolgt entsprechend der Schüler-ID auf den Umschlägen. Die Schüler*innen werden beim Start der Testsitzung gebeten, den Umschlag direkt in der Schultasche zu verstauen und zu Hause abzugeben. Die Fragebögen sollen schnellstmöglich, spätestens jedoch innerhalb einer Woche von den Erziehungsberechtigten ausgefüllt und im verschlossenen Umschlag wieder in der Schule abgegeben werden. Dort werden die verschlossenen Umschläge gesammelt und nach ca. einer Woche in einem vorfrankierten Karton an die IEA Hamburg zurückgeschickt.

Sobald alle Vorbereitungen abgeschlossen und die Arbeitsplätze eingerichtet sind, geben die Testleiter*innen die ausgedruckten Schülerauswahllisten an die Schulkoordinator*innen zurück. Die Schüler*innen werden nun in die Testräume gebeten. Zu Beginn der Testsitzung suchen sich die Schüler*innen ihren Arbeitsplatz, entfernen den Namenszettel vom Log-in-Blatt und kleben ihn vor sich auf den Tisch. Die Schüler*innenauswahllisten verbleiben an der Schule und werden dort später vernichtet. Am Ende der Testsitzung werden auch die Namenszettel von den Tischen entfernt und an die Schulkoordinator*in zurückgegeben. Somit ist das datenschutzrechtliche Prinzip der pseudonymisierten Erhebung umgesetzt, sodass keine Namen die Schule verlassen.

Alle Schüler*innen beginnen gleichzeitig unter Anleitung der Testleitung mit der Bearbeitung der PISA-Aufgaben. Die Testsitzung besteht neben den Einführungsteilen in die einzelnen Test-Abschnitte aus zwei jeweils 60-minütigen Kompetenztests sowie einem Schüler*innenfragebogen, für den eine Ausfüllzeit von ca. 50 Minuten vorgesehen ist. Zwischen den beiden Testblöcken und der Bearbeitung des Fragebogens liegt jeweils eine Pause. Um sicherzustellen, dass die Testsitzung in allen Schulen standardisiert abläuft, folgen die Testleiter*innen genau den Instruktionen im Testleiterskript, lesen Anweisungen wortwörtlich vom Skript ab und achten auf die Einhaltung der vorgegebenen Bearbeitungszeiten. Während der Testsitzung tragen die Testleiter*innen den Teilnahmestatus für jede*n Schüler*in pro Testteil in die Schüler*innenanwesenheitsliste ein. Außerdem füllen sie ein standardisiertes Testsitzungsprotokoll aus, in dem jegliche Auffälligkeiten, besondere Vorkommnisse oder Probleme während der Testsitzung dokumentiert werden. Sobald die vorgesehene Bearbeitungszeit des letzten Testteils (Schülerfragebogen) abgelaufen ist, wird die Testsitzung durch die Testleitung beendet.

Bevor die Schüler*innen am Ende der Testsitzung entlassen werden, überprüft die Testleitung, dass sich noch alle Testinstrumente (Log-in-Blätter, USB-Sticks) im Raum befinden. Alle Materialien werden durch die Testleiter*innen eingesammelt und per gesichertem Versand an die IEA Hamburg zurückgeschickt.

Im Anschluss an jede Testsitzung hält die Testleitung noch einmal kurz Rücksprache mit der Schulkoordination, um eventuelle Fragen zu klären. Da eine erfolgreiche Durchführung der PISA-Studie von einer möglichst hohen Teilnahmequote abhängt, errechnen die Testleiter*innen nach der letzten regulären Testsitzung an einer Schule die Anwesenheitsquote und informieren die Schule darüber, ob die Durchführung eines Nachtests erforderlich ist. Sollte ein Nachtest erforderlich sein, so sind die Schüler*innen aller Testgruppen an einer Schule, die im Rahmen der regulären Testsitzungen nicht anwesend waren, für die Teilnahme an einem gemeinsamen Nachtest vorgesehen. Die Durchführung des Nachtests folgt dem Ablauf einer regulären Testsitzung.

12.6.3 Datenmanagement

Für die Umsetzung der PISA-Studie gibt die OECD strukturelle Vorgaben zu Organisation der nationalen Projektzentren vor. Neben dem *nationalen Projektmanagement*- (NPM) ist dabei organisatorisch ein *nationales Datenmanagement Team* (NDM) vorgesehen, welches für die termingerechte Bearbeitung aller anfallenden Prozesse des Datenmanagements verantwortlich ist (vgl. OECD. 2020c). In Deutschland arbeiten neben dem Datenmanagement in der nationalen Projektleitung am *Zentrum für internationale Bildungsvergleichsstudien* (ZIB) an der *TUM School of Social Sciences and Technology der Technischen Universität München* noch Mitarbeiter*innen der *IEA Hamburg* aus den Bereichen Datenmanagement sowie der Kodierabteilung an den Prozessen des Datenmanagements.

Für jede PISA-Erhebung finden in einem Zeitraum über drei Jahre zwei Treffen der NPMs und NDMs statt, in welchen unter anderem spezielle Trainings für die Stichprobenziehung, Datensammlung sowie Kodierung, Datenaufbereitung und Validierung durchgeführt werden, um die hohen Standards der PISA-Studie in jedem teilnehmenden Staat einhalten zu können.

Noch bevor die eigentlichen Testungen in den PISA-Schulen durchgeführt werden, wird die internationale Datenmaske in einer speziellen Datenverwaltungssoftware, dem *IEA Data Management Expert* (DME), um nationale Adaptionen und Ergänzungen der Fragebögen angepasst. Nationale Adaptionen bezeichnen dabei nationale Änderungen in Antwortformaten, welche eindeutig einer internationalen Kategorie zugeordnet werden können. Ein Beispiel dafür sind die unterschiedlichen Schularten oder Bildungsabschlüsse (der Eltern) in den teilnehmenden Staaten. Diese Unterschiede werden nach nationaler Erfassung nach der International Standard Classification of Education (ISCED-11) entsprechend in internationalen vergleichbaren Aufschlüsselungen rekodiert (OECD et al., 2015). Diesen Schritt bezeichnet man als *Harmonisierung* der nationalen Datenmaske an ihr internationales Pendant. Die einzelnen Frageinhalte werden in ihrer (Daten-)Struktur nicht verändert, so dass hier keine nationalen Anpassungen notwendig sind. Neben den (über ggf. notwendige Adaptionen) international vergleichbar erhobenen Variablen in den resultierenden Daten besteht in jeder PISA-Runde die Möglichkeit, im Rahmen spezifischer Fragestellungen in den Hintergrundfragebögen eigene Fragen oder Fragebogenskalen zu ergänzen. Zu solchen *nationalen Ergänzungen* liegen dann nur für den jeweiligen Staat Daten vor, sodass diese nicht international angepasst beziehungsweise rekodiert werden müssen. Die Skalenhandbücher der jeweiligen PISA-Testungen in Deutschland enthalten entsprechende Übersichten zu diesen nationalen Ergänzungen (vgl. Mang et al., 2021a).

Diese hier beschriebenen Adaptionen und Harmonisierungen sowie nationalen Ergänzungen beziehen sich nur auf die sich der Kompetenztestung anschließenden Fragebögen für die Jugendlichen, die Schulleiter*innen, die Lehrkräfte und die papierbasierten Fragebögen für die Erziehungsberechtigten. Nach dem Abschluss dieser Prozeduren liegt demnach noch vor dem eigentlichen Testzeitraum eine an die deutschen Testinstrumente angepasste Datenmaske vor.

Nach der Beendigung der PISA-Testsitzungen in den Schulen, der Befragung der Lehrkräfte und Erziehungsberechtigten werden die Daten aus den Test- und Fragebogenantworten in einem streng gesicherten Verfahren in das vom internationalen Konsortium zur Verfügung gestellte Datenbanksystem (DME) eingelesen. Die DME-Software ist eine leistungsstarke, eigenständige Anwendung, die auf den meisten Windows-Systemen installiert werden kann und keine Verbindung zum Internet benötigt. Gearbeitet wird auf einer separaten Datenbankdatei mit klar definierten Datensätzen, die den verschiedenen Test- und Fragebogeninstrumenten der Studie zugeordnet sind. Der Vorteil der Software liegt auch im Auslesen der Systemdateien, welche neben den Antworten digital erfasst werden und zum Beispiel die Bearbeitungszeit einer Schülerin oder eines Schülers enthalten.

Eine Vielzahl von Datenbereinigungsschritten und Überprüfungen in Form sogenannter Konsistenz- und Validitätskontrollen folgen in einem Zeitraum von circa drei Monaten sowohl auf Seiten des nationalen Datenmanagementteams als auch auf Seiten der internationalen Vertragsnehmer. Absprachen erfolgen immer im engen und regelmäßigen Austausch beider Parteien. Die internationalen Vertragsnehmer stellen im Spätsommer des Jahres der Berichtslegung die aufbereiteten und mit Schätzwerten der Kompetenzen und weiteren Skalen versehenen Datensätze dem nationalen Projektteam zur internen Nutzung zur Verfügung. Neben diesen Datensätzen gibt es weitere Berichte mit Informationen zur Qualitätssicherung sowie Validierung. Erst mit dem Stichtag der Veröffentlichung der Berichtserstattung werden die internationalen Daten aller beteiligten Staaten über die Webseite der OECD zur freien Nutzung zur Verfügung gestellt. Die nationalen, deutschen Daten werden nach Berichterstattung über das *Forschungsdatenzentrum* (FDZ) des *Instituts für Qualitätsentwicklung im Bildungswesen* (IQB) den deutschsprachigen wissenschaftlichen Nutzern freigegeben (vgl. Mang et al., 2021b).

Literatur

Mang, J., Seidl, L., Schiepe-Tiska, A., Tupac-Yupanqui, A., Ziernwald, L., Doroganova, A., Weis, M., Diedrich, J., Heine, J.-H., González Rodríguez, E. & Reiss, K. (2021a). *PISA 2018 Skalenhandbuch. Dokumentation der Erhebungsinstrumente.* Waxmann. https://doi.org/10.31244/9783830994961

Mang, J., Heine, J.-H., Weis, M., Diedrich, J., Schiepe-Tiska, A., Ziernwald, L., Tupac-Yupanqui, A., Doroganova, A., González Rodríguez, E., Reiss, K., Klieme, E. & Köller, O. (2021b). *Programme for International Student Assessment 2018 (PISA 2018)* Version 1. [Datensatz]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.5159/IQB_PISA_2018_v1

OECD. (2020c). *PISA national project manager manual.* https://www.oecd.org/pisa/pisaproducts/PISA-2022-National-Project-Manager-NPM-Manual.pdf

OECD, Eurostat & UNESCO Institute for Statistics. (2015). *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*. OECD Publishing. https://doi.org/10.1787/9789264228368-enhttps://doi.org/10.1787/9789264228368-en

12.7 Methoden zu Skalierungen und Klassifikationsanalysen für einzelne Kapitel des Berichtsbands 2022

Jörg-Henrik Heine & Sabine Patzl

Für einzelne Kapitel in diesem Berichtsband für PISA 2022 wurden für vertiefende Analysen einzelne Skalen der Hintergrundfragebögen neu gebildet (skaliert). Diese Skalierungen wurden notwendig, weil sich die Fragen (bzw. Item-)Zusammensetzung in den von der OECD und deren Vertragsnehmer (ETS) bereitgestellten skalierten und abgeleiteten Variablen für die angestrebten Vergleiche zwischen unterschiedlichen PISA-Erhebungsrunden verändert haben.

Daneben wurde für das Kapitel 8 als Klassifikationsverfahren über insgesamt fünf Dimensionen des Mathematikunterrichts die latente Profilanalyse (LPA) angewendet, um aus einer personenzentrierten Perspektive typische Wahrnehmungsmuster zum Mathematikunterricht aus Sicht der Jugendlichen zu identifizieren. In diesem Abschnitt werden die methodischen Hintergründe der eingesetzten Verfahren knapp erläutert und deren Anwendung dokumentiert.

12.7.1 Skalierungen für abgeleitete Variablen in den Fragebögen

Um Veränderungen in einzelnen abgeleiteten Variablen aus dem Fragebogen für die Jugendlichen über verschiedene PISA-Erhebungsrunden zu untersuchen, wurden die betreffenden psychometrischen Skalen einer marginalen Trendschätzung unterzogen (vgl. Gebhardt & Adams, 2007). Bei dieser Trendschätzung basiert der Vergleich über die verschiedenen Zeitpunkte hinweg lediglich auf jeweils unveränderten (Link-)Items. Das bedeutet, dass nur jene Fragen aus dem Hintergrundfragebogen berücksichtigt werden, welche zu allen relevanten Testzeitpunkten in äquivalenter Weise eingesetzt wurden (Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019, 2022). Im Rahmen einer gleichzeitigen Kalibrierung (eng. *concurrent calibration*, vgl. Dorans et al., 2007; Kolen & Brennan, 2014; Winkersky & Lord, 1984) werden bei der Skalierung die betreffenden abgeleiteten Variablen (psychometrischen Skalen) basierend auf den einzelnen Fragen in den Fragebogen neu gebildet. Diese Skalierungen erfolgten für alle Skalen grundsätzlich in drei Schritten.

Zunächst werden im ersten Schritt die mit ordinal aufsteigenden Kardinalzahlenwerten *gescoreten* Antworten der Jugendlichen auf den fest vorgegebenen vier- bis fünfstufigen Antwortskalen aus den jeweiligen Datensätzen der entsprechenden PISA-Erhebungsrunden identifiziert. Items, deren Antwortskala im Sinne des zu erfassenden Merkmals negativ formuliert sind, wurden entsprechend umkodiert. Für die Skalierungen in Kapitel 4 wurden neben den Daten für Deutschland aus der aktuellen PISA-Erhe-

bungsrunde 2022 auch die über das Forschungsdatenzentrum (FDZ) verfügbaren Daten zu PISA 2012 (vgl. Prenzel et al, 2015) und 2003 (Prenzel et al., 2007) in Deutschland herangezogen. Bei den Skalierungen für Kapitel 8 wurden die internationalen Daten der PISA-Erhebungsrunden 2022 und 2012 verwendet, wobei diese auf die OECD-Staaten reduziert wurden. Für Kapitel 10 wurde lediglich eine Skala bestehend aus vier Items aus den PISA-Daten der aktuellen Erhebungsrunde zu „*Gefühlen zum Selbstlernen*" (vgl. Kapitel 10, Tabelle 10.14) neu skaliert.

Im zweiten Schritt erfolgte die Kalibrierung und die Verlinkung der jeweiligen PISA-Erhebungsrunden (vgl. Winkersky & Lord, 1984). Dabei wurden die Antworten der Jugendlichen je nach Datengrundgrundlage aus den jeweiligen PISA-Erhebungsrunden (2003, 2012 und 2022) zu einer Datei kombiniert. Für die ausgewählten Fragebogen-Items der einzelnen Skalen wurden mit dieser neu zusammengestellten Datenbasis über alle enthaltenen Messzeitpunkte gemeinsame neue Itemparameter bestimmt. Die Itemkalibrierungen erfolgten mit dem *R*-Paket pairwise (Heine, 2023a) für die freie Statistikumgebung *R* (*R* Core Team, 2023).

Im dritten Schritt wurde auf Basis der ermittelten und fixierten Itemparameter die Personenparameter der Jugendlichen für die jeweiligen PISA-Erhebungsrunden (2003, 2012 und 2022) bestimmt. Die Skalierung der Personenparameter erfolgte ebenfalls mit dem *R*-Paket pairwise (vgl. Heine, 2023a), wobei dafür sogenannte *weighted likelihood estimates* (WLE; Warm, 1989) als Messwerte für die jeweiligen Einstellungen oder Unterrichtswahrnehmungen (vgl. z. B. Kapitel 8) der Jugendlichen bestimmt wurden.

Für Kapitel 8 wurden die zunächst auf der Logit-Metrik vorliegenden Personenparameter aus der IRT-Skalierung für die Skala FAMCON (*Vertrautheit mit mathematischen Inhalten*) einer *z*-Transformation unterzogen. Dadurch wurde der Mittelwert über alle OECD-Staaten für die betrachteten Messzeitpunkte so verschoben, dass in der gewählten Referenz-Erhebungswelle der OECD-Skalenmittelwerte bei $M = 0$ mit einer Standardabweichung von $SD = 1$ liegt. Zur Darstellung von Veränderungen in der Wahrnehmung des Mathematikunterrichts durch die Fünfzehnjährigen zwischen den PISA-Erhebungsrunden 2012 und 2022 wurden für das Kapitel 8 so insgesamt die vier Skalen mit den internationalen Variablenbezeichnungen FAMCON (*Vertrautheit mit mathematischen Inhalten*), TEACHSUP (*Unterstützung durch die Lehrkraft*), EXPOFA (*Häufigkeit innermathematischer und einfacher Anwendungsaufgaben*) und DISCLIM (*Disziplin im Klassenzimmer*) einer Reskalierung, wie oben beschrieben, unterzogen. Im Online-Anhang von Kapitel 8 sind in der Tabelle 8.2web („Skalenzusammensetzung für die Trendberechnungen") unter den entsprechenden inhaltlichen Skalenbeschreibungen die einzelnen Items mit ihren Frageinhalten aus den beiden PISA-Erhebungsrunden 2012 und 2022 aufgeführt. Für Kapitel 4 wurden die Skalen zu Einstellungen gegenüber der Mathematik mit der internationalen Variablenbezeichnungen MATHEFF (*Selbstwirksamkeitserwartung bezüglich innermathematischer und einfacher Anwendungsaufgaben*), INTMAT (*Freude und Interesse an Mathematik*), ANXMATH (*Ängstlichkeit im Fach Mathematik*) sowie INSTMOT (*mathematikbezogene instrumentelle Motivation*) einer Reskalierung über die PISA-Erhebungsrunden 2003, 2012 und 2022 unterzogen. In der Abbil-

dung 4.2web ist im Online-Anhang zu Kapitel 4 eine Übersicht zu den für die jeweiligen Skalierungen berücksichtigten Items gegeben. Die Items der für Kapitel 10 neu gebildeten Skala sind in Tabelle 10.14 des Kapitels 10 dokumentiert.

12.7.2 Klassifikationsanalysen zu typischen Wahrnehmungsmustern des Mathematikunterrichts

Zur Identifikation von typischen Wahrnehmungsmustern zum Mathematikunterricht aus Sicht der Jugendlichen wurde für das Kapitel 8 dieses Berichtsbandes die *Latent Profile Analysis* (LPA; Gibson, 1959) eingesetzt. Als Klassifikationsverfahren ermöglicht die Anwendung der LPA über insgesamt fünf Dimensionen des Mathematikunterrichts aus einer personenzentrierten Perspektive die Identifikation von Gruppen von Jugendlichen mit einer differentiellen Wahrnehmung des Mathematikunterrichts. Die LPA stellt eine Generalisierung (Gibson, 1959) des latenten Strukturmodells von Lazarsfeld, (1950; 1959) dar, welches auch die theoretische Grundlage für die *Latent Class Analysis* (LCA; Lazarsfeld, 1950, Formann, 1984) darstellt. Das übergreifende Konzept des latenten Strukturmodells, das diesen beiden statistischen Modellen (LCA und LPA) zugrunde liegt, besteht darin eine Population nicht als homogene Gruppe zu betrachten. Stattdessen wird eine unterschiedliche Anzahl von Gruppen (Sub-Populationen, *Klassen*) angenommen, innerhalb derer sich die Personen hinsichtlich ausgewählter Merkmale ähnlicher sind (typische Muster bzw. Merkmalskonfigurationen) als Personen zwischen den Klassen (Gruppen). Im Gegensatz zur LCA, welche auf die Analyse von kategorialen beziehungsweise nominal skalierten Merkmalsvariablen ausgerichtet ist, dient die LPA der Identifikation von typischen Musterkonfigurationen basierend auf metrischen beziehungsweise kontinuierlichen Variablen (Gibson, 1959). Als kontinuierliche Messgrößen zur Klassifikation wurden in den hier durchgeführten Analysen die von dem internationalen Vertragsnehmer *Educational Testing Service* (ETS) bereitgestellten abgeleiteten Skalen für *Disziplin im Klassenzimmer* (DISCLIM), *Unterstützung durch die Lehrkraft* (TEACHSUP), *Mathematisches Argumentieren* (COGACRCO), *Ermutigung zum mathematischen Denken* (COGACMCO) sowie *Häufigkeit innermathematischer und einfacher Anwendungsaufgaben* (*EXPOFA*) berücksichtigt. Sämtliche Berechnungen und Modellschätzungen wurden in der freien Statistikumgebung *R* (R Core Team, 2023) mit dem „Wraper“ *R*-Paket tidyLPA (Rosenberg et al., 2021) durchgeführt. Zur Modellschätzung kann dieses *R*-Paket entweder auf die kommerzielle Software *Mplus* (Muthén & Muthén, 1998–2017) oder auf das *R*-Paket mclust (Fraley et al., 2022) zurückgreifen. Die in Kapitel 8 berichteten Analysen wurden unter Rückgriff auf das *R*-Paket mclust durchgeführt. Es ist wichtig zu betonen, dass alle Modellschätzungen für die Klassifikation zunächst anhand ungewichteter Stichprobendaten durchgeführt wurden. Im direkten Vergleich zur Verwendung von gewichteten Stichprobendaten für Klassifikationsanalysen zeigt sich typischerweise, dass zumindest bei einfachen (beschreibenden) Klassifikationsmodellen – wie dem hier angewendeten Modell – keine bedeutsamen Unterschiede in der

Personenklassifikation aus der Analyse ungewichteter im Vergleich zu gewichteten Stichprobendaten resultieren (Heine et al., in Druck; Heine, 2023b). Die mit mclust geschätzten (LPA-)Modelle können zu deren Identifikation mit unterschiedlichen Restriktionen spezifiziert werden. Im Kern geht es dabei um die Entscheidung, ob die Varianzen und / oder Kovarianzen der klassifizierenden Variablen restringiert oder frei geschätzt werden sollen. Konkret können über tidyLPA mit mclust folgende Modelle (Kombinationen) spezifiziert werden:

1) Gleiche Varianzen und Kovarianzen fixiert auf 0 (Modell 1)
2) Frei geschätzte Varianzen und Kovarianzen fixiert auf 0 (Modell 2)
3) Gleiche Varianzen und gleichbleibende Kovarianzen (Modell 3)
4) Frei geschätzte Varianzen und frei geschätzte Kovarianzen (Modell 6)

Mithilfe eines analytischen Hierarchieprozesses (AHP; vgl. Saaty, 1987) und unter Hinzuziehung unterschiedlicher, Kullback-Leibler-Distanz basierter (vgl. Kullback & Leibler, 1951), informationstheoretischer Kriterien zum relativen Modellvergleich, kann aus den geschätzten Modellen das „relativ am besten passende Modell" bestimmt werden. Ziel ist es, dasjenige Modell zu identifizieren, welches innerhalb einer Kette der konvergierenden Klassen-Modelle den geringsten Wert für ein informationstheoretisches Kriterium aufweist.

Für die hier dokumentierten Analysen wurden für alle Modellspezifikationen (1, 2, 3 und 6) jeweils explorativ Modelle unter der Annahme von jeweils 1 bis 9 latenten Klassen (Gruppen / Sub-Populationen) geschätzt. Für den relativen Modellvergleich werden die beiden informationstheoretischen Kriterien AIC (Akaike, 1974) und BIC (Schwarz, 1978) berichtet. Danach ergeben sich für die aktuelle Erhebungsrunde in Deutschland die in Tabelle 12.7.1 dargestellten Ergebnisse zum relativen Modellvergleich.

Tabelle 12.7.1: Informationstheoretische Kriterien zur relativen Modellwahl

Modell	Anzahl Klassen	AIC	BIC
1	1	103282	103362
1	2	100852	100979
1	3	99774	99949
1	4	**99361**	**99583**
1	5	99368	99636
1	6	98941	99257
1	7	98864	99227
1	8	-	-
1	9	-	-
6	1	99470	99651
6	2	97241	97610
6	3	96827	**97385**
6	4	**96749**	97495
6	5	-	-
6	6	96526	97649
6	7	-	-
6	8	96262	97760
6	9	-	-

Anmerkungen: Modell 1: Restriktion auf gleiche Varianzen und auf null festgelegte Kovarianzen innerhalb der latenten Klassen, Modell 6 (keine Restriktionen): Variierende Varianzen und variierende Kovarianzen innerhalb der latenten Klassen, Modelle 2 und 3: nicht konvergiert für 1–9 latente Klassen.

Die Inspektion der in Tabelle 12.7.1 gegebenen Werte zeigt, dass nicht alle Modellspezifikationen zu einem konvergierenden Modell führen. So konnten insbesondere für die Modellspezifikationen 2 und 3 keine konvergierenden Modelle gefunden werden. Für die Modellspezifikationen 1 und 6 führten die Annahmen bestimmter Klassenanzahlen zu nicht konvergierenden Modellen (vgl. Tabelle 12.7.1; Modelle ohne Werte für AIC oder BIC). Um auf dieser Basis ein für die inhaltliche Interpretation geeignetes Modell auszuwählen, müssen aus einer methodischen Perspektive folgende Bedingungen erfüllt sein:

1) Das Modell muss konvergieren; das heißt die Modellparameter (hier die Klassenzuordnungswahrscheinlichkeiten) müssen innerhalb der vorgegebenen Anzahl von Iterationen zur Schätzung mit einer vorgegebenen Genauigkeit bestimmbar sein.
2) Das bevorzugte Modell muss ein nachfolgendes, konvergierendes Modell mit einer höheren Klassenanzahl aufweisen.

Nach diesen beiden Kriterien ist zunächst zu berücksichtigen, dass nicht alle Modelle aus den spezifizierten Modellen konvergieren. Legt man das informationstheoretische Kriterium BIC (und AIC) zugrunde und betrachtet die Kette der konvergierenden Klassen-Modelle mit Kovarianz- und Varianzrestriktion (Modellspezifikation 1 mit 1–9 Klas-

sen), so weist das Klassen-Modell mit vier latenten Klassen den geringsten Wert sowohl für den BIC ($BIC_{6.4}$ = 99583) als auch für den AIC ($AIC_{6.4}$ = 99361) auf. Das entsprechende Modell mit fünf latenten Klassen konvergiert zwar ebenfalls, weist aber höhere Werte für den BIC und AIC auf.

Für die Modellspezifikation 6 (frei geschätzte Varianzen und Kovarianzen) wird nach dem informationstheoretischen Kriterium BIC (Schwarz, 1978), welches ein stärkeres Gewicht auf die Sparsamkeit (Einfachheit) bei der vergleichenden Modellwahl legt (vgl. Rost, 2004), das Modell mit drei latenten Klassen bevorzugt. Aus den Ergebnissen der Modellschätzungen für beide Modellspezifikationen (1 und 6) wurde jeweils eine (kategoriale) Klassifikationsvariable bestimmt (einmal 4-stufig und einmal 3-stufig). Diese Klassifikationsvariablen weisen die Jugendlichen aus der PISA-Stichprobe für Deutschland jeweils einer der latenten Klassen mit maximaler Wahrscheinlichkeit zu. Für die in Kapitel 8 berichteten Analysen wurde unter Einbezug von inhaltlich interpretatorischen Gesichtspunkten der jeweiligen Profile für die fünf Merkmalsdimensionen des Mathematikunterrichtes das Klassifikationsergebnis auf der Grundlage des Modelles mit vier latenten Klassen dargestellt und untersucht.

Die gewichtete, mittlere maximale Klassenzuordnung als Maß für die Reliabilität des Klassifikationsergebnisses beträgt für dieses Modell $r = 0.79$, was ein angemessener Wert für die Zuverlässigkeit der Klassifikation ist. Für jede dieser vier latenten Klassen wurden Mittelwerte für die klassifizierenden, metrischen Variablen (DISCLIM, TEACHSUP, COGACRCO, COGACMCO, EXPOFA) bestimmt, welche die Grundlage für die in Kapitel 8 in der Abbildung 8.4 gegebenen Profildarstellungen sind. Die entsprechenden Berechnungen zu den Klassengrößen und den Klassen-Mittelwerten basieren auf gewichteten Stichprobendaten unter Berücksichtigung der PISA-typischen Methodologie (vgl. Heine & Reiss, 2019), wobei zur Berechnung der Koeffizienten die Stichprobengewichte und deren 80 Replikatgewichte verwendet wurden. Für diese Analysen und Berechnungen wurde einerseits das *R*-Paket BIFIEsurvey (Robitzsch et al., 2022) und andererseits die Software SPSS in Verbindung mit den offiziellen Macro-Skripten aus dem IEA-IDB-Analyzer eingesetzt. Die inhaltlich orientierten Interpretationen der Profile der latenten Klassen aus Merkmalen zur Unterrichtswahrnehmung für den Mathematikunterricht finden sich in den entsprechenden Abschnitten in Kapitel 8 dieses Berichtsbandes.

Literatur

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatik Control, AC-19* (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Dorans, N. J., Pommerich, M., & Holland, P. W. (Hrsg.). (2007). *Linking and aligning scores and scales.* Springer. https://doi.org/10.1007/978-0-387-49771-6

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in Theorie und Anwendung.* Beltz.

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., & Fop, M. (2022). mclust: Gaussian Mixture Modelling for Model-Based clustering, classification, and density estimation (5.4.10) [Software]. https://CRAN.R-project.org/package=mclust

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*(3), 229–252. https://doi.org/10.1007/BF02289845

Heine, J.-H. (2023a). *pairwise: Rasch Model parameters by pairwise algorithm* (Version 0.6.1-0) [Software]. 17. 4. 2023 https://CRAN.R-project.org/package=pairwise

Heine, J.-H. (2023b). *Stichprobengewichtung bei der Personenzentrierten Analyse von Konfigurationen aus kategorialen Daten*. Vortrag auf der 16. Tagung der Fachgruppe Methoden & Evaluation der DGPs, Konstanz, 11.09.2023–15.09.2023.

Heine, J.-H., & Hartmann, F. G., & Tarnai, Ch., (in Druck). Exploring temporal pattern of intergenerational educational mobility in Germany using weighted Prediction Configural Frequency Analysis. In M. Stemmler, W. Wiedermann & F. Huang (Hrsg.) *Dependent data in Social Science research* (2. Aufl.). Springer.

Heine, J.-H., & Reiss, K. (2019). PISA 2018 – die Methodologie. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018 Grundbildung im internationalen Vergleich* (S. 241–58). Waxmann.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking.* Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22* (1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lazarsfeld, P. F. (1950). The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Hrsg.), *Studies in social psychology in World War II: Bd. IV Measurement and prediction* (S. 362–412). Princeton University Press.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Hrsg.), *Formulations of the person and the social context* (Bd. 3, S. 476–543). Mcgraw-Hill. http://archive.org/details/psychologyastudy017916mbp

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8. Aufl.). Muthén & Muthén.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J. & Schiefele, U. (2007). Programme for International Student Assessment 2003 (PISA 2003) (Version 1) [Datensatz]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2003_v1

Prenzel, M., Sälzer, C., Klieme, E., Köller, O., Mang, J., Heine, J.-H., Schiepe-Tiska, A. & Müller, K. (2015). Programme for International Student Assessment 2012 (PISA 2012) (Version 5) [Datensatz]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2012_v5

R Core Team. (2023). R: A language and environment for statistical computing [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice, 26*(4), 444–465. https://doi.org/10.1080/0969594X.2018.1433633

Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences*, 4. https://doi.org/10.1186/s42409-022-00039-w

Robitzsch, A., BIFIE, & Oberwimmer, K. (2022). BIFIEsurvey: Tools for survey statistics in educational assessment (Version 3.4-15) [Software]. http://CRAN.R-project.org/package=BIFIEsurvey

Rosenberg, J. M., Lissa, C. van, Schmidt, J. A., Beymer, P. N., Anderson, D., & Schell, M. J. (2021). tidyLPA: Easily carry out Latent Profile Analysis (LPA) using open-source or commercial software (1.1.0) [Software]. https://CRAN.R-project.org/package=tidyLPA

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., vollst. überarb. u. erw. Aufl.). Huber.

Saaty, R. W. (1987). The analytic hierarchy process – What it is and how it is used. *Mathematical Modelling, 9*(3), 161–176. https://doi.org/10.1016/0270-0255(87)90473-8

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Winkersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement,* 8, 347–364. http://dx.doi.org/10.1002/j.2330-8516.1983.tb00028.x

12.8 Vertiefende Trendanalysen für PISA 2018 bis 2022

Alexander Robitzsch & Oliver Lüdtke

In diesem Abschnitt werden vertiefende Trendanalysen für Deutschland berichtet. Diese Analysen untersuchen, ob sich die offiziell berichteten Trends auch bei einer Variation des Analysevorgehens als belastbar erweisen. Konkret werden marginale und bedingte Trendschätzungen der offiziellen originalen Trendschätzung gegenübergestellt. Außerdem wird der Einfluss der Behandlung fehlender Item-Antworten und der Implementation des adaptiven Testdesigns hinsichtlich des deutschen Trends von PISA 2018 nach PISA 2022 studiert.

12.8.1 Marginale Trendschätzung

Die in den internationalen und nationalen Berichtsbänden veröffentlichten Trends für Deutschland beruhen auf internationalen Skalierungen und werden als originale Trends bezeichnet. Alternativ können jedoch auch marginale Trends für Deutschland berechnet werden, die auf Basis einer nationalen Skalierung ermittelt werden (Robitzsch et al., 2017). Marginale Trendschätzungen besitzen den Vorteil, weniger stark von länderspezifischem differentiellen Item-Funktionieren (Differential Item Functioning) betroffen zu sein (Robitzsch & Lüdtke, 2019).

In diesem Abschnitt wird der marginale Trend mit Hilfe eines Zweigruppen-Generalized-Partial-Credit-Modells unter der Annahme invarianter Item-Parameter ermittelt. Die beiden Gruppen entsprechen den PISA-Erhebungen 2018 und 2022 für deutsche Schülerinnen und Schüler in den beiden PISA-Studien. Das verwendete Item-Response-Modell stimmt also mit dem in der internationalen Skalierung verwendeten IRT-Modell überein. Allerdings sind die Item-Parameter national für Deutschland ermittelt und es ist kein Rückgriff auf eine internationale Metrik notwendig. Die Ergebnisse wurden so linear transformiert, dass der Mittelwert und die Standardabweichung mit den offiziell berichteten Werten in PISA 2018 übereinstimmen. Die marginalen Skalierungen wurden eindimensional separat für die drei Domänen Mathematik, Lesen und Naturwissenschaften durchgeführt.

Tabelle 12.8.1 berichtet die originale und marginale Trendschätzung für die deutschen Mittelwerte für den Trend von PISA 2018 zu PISA 2022. Die marginalen Trendschätzungen fallen relativ ähnlich zu den originalen Trendschätzungen aus. Interessanterweise ist der original berichtete Leistungsabfall in Mathematik mit -25.4 Punkten statistisch signifikant größer als der originale Trend in Lesen mit -18.4 Punkten. Die marginalen Trends für Mathematik und Lesen stimmten jedoch mit -22.0 und -22.4 Punkten praktisch überein. Auch im marginalen Trend zeigte sich, dass der Leistungsabfall in Naturwissenschaften mit -10.9 Punkten am geringsten in den drei Domänen ausgeprägt war.

Tabelle 12.8.1: Originale und marginale Trendschätzung für Mittelwerte

Domäne	original	marginal
Mathematik	−25.4	−22.0
Lesen	−18.4	−22.4
Naturwissenschaften	−10.7	−10.9

12.8.2 Bedingte Trendschätzung

In diesem Abschnitt stellen wir dem originalen Trend einem bedingten Trend (vgl. Robitzsch et al., 2020) gegenüber, der die hypothetischen Leistungsveränderungen berichtet, wenn sich die demografische Zusammensetzung der Schülerschaft von PISA 2018 nach PISA 2022 nicht geändert hätte. Es werden nur die zwei relevantesten Kovariaten berücksichtigt: der Migrationshintergrund und der Sozialstatus HISEI. Der Anteil von Schüler*innen mit Migrationshintergrund stieg von PISA 2018 (22.2 %) nach PISA 2022 (25.5 %), während sich die HISEI-Mittelwerte praktisch nicht unterschieden (M = 51.8 in PISA 2018 sowie M = 51.9 in PISA 2022).

Die bedingten Trends wurden auf Basis der von der OECD herausgegebenen Plausible Values bestimmt. Fehlende Werte in den beiden Kovariaten wurden mit einem Mehrebenenimputationsmodell (Level 1: Schüler, Level 2: Schulen) unter *fully conditional specification* separat für beide PISA-Erhebungen imputiert. Die bedingten Trends wurden mit der Methode des Regression Matching (Schafer & Kang, 2008) berechnet. Dazu wurde eine lineare Regression der jeweiligen Kompetenzdomäne auf Terme der Kovariaten in PISA 2022 ermittelt und die Kovariatenausprägungen der Schüler*innen aus PISA 2018 zur Bestimmung vorhergesagter Kompetenzwerte benutzt. Der gewichtete Mittelwert aus diesen Vorhersagen ergibt dann die Mittelwertschätzung für PISA 2022, wenn sich die demografische Zusammensetzung der Schülerschaft nicht geändert hätte. Damit lässt sich die Trendschätzung direkt berechnen. Es wurden zwei Spezifikationen für die bedingte Trendschätzung implementiert. Das erste Modell (bed. L1) nutzt nur die Level-1-Kovariaten Migrationshintergrund und HISEI, wobei auch deren Interaktion sowie der quadratische Term von HISEI in die Regression einflossen. Im zweiten Modell (bed. L1+L2) gingen auch Schulmittelwerte (aggregierte Werte der Level-1-Kovariaten auf Level 2 der Schule) des Migrationshintergrundes und des HISEI ein. Damit wurden auch Effekte einer sich möglicherweise verschlechternden Komposition in Schulen berücksichtigt.

In Tabelle 12.8.2 wird der originale Trend den beiden bedingten Trends gegenübergestellt. Die bedingten Trends zeigten erwartungsgemäß geringere Leistungsabfälle, konnten jedoch die stark negativen Trends nur zu einem geringen Teil erklären. Es soll hervorgehoben werden, dass die Leistungsabfälle in den bedingten Trends geringer ausgeprägt waren, wenn zusätzlich für Effekte der Komposition kontrolliert wurde. Insgesamt waren die Leistungsabfälle der bedingten Trends auf Basis von Level-1- und

Level-2-Kovariaten zwischen 2.4 bis 3.7 Punkten kleiner als die originalen Trends. Sich ändernde Schülerzusammensetzungen in der deutschen Population können also die Negativtrends von PISA 2018 nach PISA 2022 nur zu einem kleinen Teil aufklären.

Tabelle 12.8.2: Originale und bedingte Trendschätzungen für Mittelwerte

Domäne	original	bed. L1	bed. L1+L2
Mathematik	−25.4	−23.6	−22.5
Lesen	−18.4	−16.6	−15.9
Naturwissenschaften	−10.7	−8.5	−7.0

Anmerkung: bed. L1 = bedingte Trendschätzung mit Level-1-Kovariaten; bed. L1+L2 = bedingte Trendschätzung mit Level-1-Kovariaten und aggregierten Kovariaten auf Level-2.

12.8.3 Konsequenzen der Behandlung fehlender Item-Antworten

Es könnte vermutet werden, dass die Leistungsabfälle in PISA 2022 Ursache nicht-engagierten Item-Antwortverhaltens der Schüler*innen sein könnte. Nichtengagiertes Antwortverhalten kann sich in fehlenden Item-Antworten niederschlagen. In diesem Abschnitt wurden die marginalen Trendschätzungen zusätzlich berechnet, indem fehlende Item-Antworten als nicht administriert (d.h. als Missing) behandelt wurden. Die Behandlung fehlender Item-Antworten als falsch (siehe Abschnitt 12.8.1) und als Missing können dabei als zwei Extrempole möglicher Analysestrategien fehlender Item-Antworten angesehen werden (Robitzsch, 2021).

Tabelle 12.8.3 führt die beiden marginalen Trendschätzungen auf. Bei der Behandlung fehlender Item-Antworten als Missing fielen die negativen marginalen Trends etwas geringer als bei der Behandlung als falsch aus. Es scheint daher ein verändertes Testverhalten unwahrscheinlich für eine Erklärung der deutlichen negativen Trends für Deutschland zu sein.

Tabelle 12.8.3: Marginale Trendschätzung mit verschiedenen Behandlungen fehlender Item-Antworten

Domäne	original	marg. 0	marg. NA
Mathematik	-25.4	-22.0	-19.0
Lesen	-18.4	-22.4	-21.5
Naturwissenschaften	-10.7	-10.9	-9.4

Anmerkung: marg. 0 = marginale Trendschätzung bei der Behandlung fehlender Item-Antworten als falsch; marg. NA = marginale Trendschätzung bei der Behandlung fehlender Item-Antworten als Missing.

12.8.4 Einfluss der adaptiven Testung

In diesem Abschnitt werden die Effekte des Wechsels zu einem adaptiven Testdesign in der Testdomäne Mathematik untersucht. Dazu wird ausgenutzt, dass für Mathematik in PISA 2022 eine experimentelle Variation der Testadministration umgesetzt wurde. Rund ein Viertel aller Schüler*innen erhielten ein lineares nicht adaptives Design, während der Rest der Schüler*innen der adaptiven Multi-Stage-Testung zugewiesen wurde.

Die mittlere Testleistung der deutschen Schüler*innen in Mathematik in PISA 2022 betrug 476.0 im linearen Design, während die mittlere Mathematikleistung im adaptiven Design 477.9 war. Die Trendschätzung für Deutschland wäre demzufolge in einem linearen Design noch negativer als im adaptiven Testdesign ausgeprägt. Allerdings war der Anteil der Schüler*innen, die mindestens ein nicht erreichtes Item (not reached item) hatte, mit 15.4 % im adaptiven Design deutlich höher gegenüber 5.2 % im linearen Design. Die leicht höhere Testleistung im adaptiven Design ist demzufolge primär eine Folge der Behandlung nicht erreichter Items in der Skalierung als Missing. Unabhängig von der Methode der Missingbehandlung kann also nicht behauptet werden, dass der Einsatz des adaptiven Testdesigns zu höheren Testleistungen führt. Obwohl wir die behaupteten Vorteile der adaptiven Testung hinsichtlich präziserer Schätzungen für fragwürdig halten (Robitzsch & Lüdtke, 2021), kann ausgeschlossen werden, dass die negativen originalen Trends in Mathematik für Deutschland primär eine Folge des adaptiven Testdesigns sind.

12.8.5 Resümee

In diesem Abschnitt wurden vertiefende Trendanalysen für die deutschen PISA-Trends von 2018 nach 2022 durchgeführt. Es zeigte sich, dass die deutlichen negativen Trends nicht oder nur zu einem sehr geringen Anteil auf methodische Effekte zurückführbar sind. Demzufolge können die berichteten Trends aus methodischer Sicht als belastbar angesehen werden und ohne größere Vorbehalte in Bildungspolitik und Öffentlichkeit interpretiert werden.

Literatur

Robitzsch, A. (2021). On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using PISA 2018 mathematics data. *European Journal of Investigation in Health, Psychology and Education, 11*(4), 1653–1687. https://doi.org/10.3390/ejihpe11040117

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice, 26*(4), 444–465. https://doi.org/10.1080/0969594X.2018.1433633

Robitzsch, A., & Lüdtke, O. (2021). Reflections on analytical choices in the scaling model for test scores in international large-scale assessment studies. *PsyArXiv*. 31 August 2021. https://doi.org/10.31234/osf.io/pkjth

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. https://doi.org/10.31244/9783830993193

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods, 13*(4), 279–313. https://doi.org/10.1037/a0014268

12.9 Fazit

Die vorliegende Dokumentation zu den Methoden der Datenerhebung und Auswertung beleuchtet die wesentlichen Aspekte, die für die reibungslose Durchführung der PISA-Studie 2022 in Deutschland erforderlich sind. Dargestellt wurden die entscheidenden Schritte, um die Qualität und Aussagekraft der gesammelten Daten sicherzustellen. Im Rahmen der PISA-Studie wurden umfangreiche Planungs- und Vorbereitungsmaßnahmen im nationalen Projektzentrum, bei der IEA Hamburg sowie in den Schulen, in denen die Erhebung stattfand, durchgeführt. Diese umfangreiche Planung und Vorbereitung und die anschließende Umsetzung der in diesem Kapitel beschriebenen Aufgabenbereiche, sowohl auf nationaler Ebene als auch in Abstimmung mit den internationalen Vertragsnehmern der OECD, stellen eine entscheidende Grundlage für die Validität und Zuverlässigkeit der erhobenen Daten dar.