PISA 2025

Samuel Greiff
Jennifer Diedrich
Tamara Kastorff
Frank Goldhammer
Olaf Köller
(Hrsg.)

Bildung im Spannungsfeld von Innovation und Transformation

KAPITEL 14

ZIB

WAXMANN

Kapitel 14: Methodische Grundlagen der PISA-Studie 2025

Philipp Sterner, Claudia Bovensiepen, Carola Bretsch, Elisabeth González Rodríguez, Tamara Kastorff, Oliver Lüdtke, Sabine Meinck, Maren Müller, Lisa Teufele, Sabrina Wagner & Jennifer Diedrich[1]

In diesem Kapitel werden die methodischen Grundlagen der PISA-Studie in der ▶ Erhebungsrunde 2025 beschrieben. Speziell werden die Erhebungsinstrumente, die ▶ Stichprobe sowie ihre Ziehung, die Testorganisation und -durchführung, das Testdesign sowie die ▶ Skalierung der Testdaten und abschließend der Umgang mit fehlenden Werten in den Daten genauer erläutert. Informationen zu PISA generell sowie Besonderheiten der Erhebungsrunde 2025 und dem deutschen Berichtsband finden sich im Kapitel *Die PISA-Studie 2025*.

In der Vorbereitung, Durchführung und Auswertung der PISA-Studie sind viele Institutionen involviert. Diese werden im Folgenden aufgelistet, um das Verständnis des Kapitels zu erhöhen. Die Rollen dieser Institutionen werden in den entsprechenden Abschnitten beschrieben.

- ZIB: Zentrum für internationale Bildungsvergleichsstudien e.V.
- IEA: International Association for the Evaluation of Educational Achievement
- OECD: Organisation für wirtschaftliche Zusammenarbeit und Entwicklung
- ACER: Australian Council for Educational Research
- cApStAn: The company of Apor, Steve and Andrea

Neben diesem Kapitel finden sich in Teil V des PISA-Berichtsbands außerdem die Online-Anhänge und das Glossar mit wichtigen Begriffen. In den Online-Anhängen werden die Ergebnisse der Analysen des Berichtsbands detailliert dargestellt; zum Beispiel finden sich dort die berücksichtigten Merkmale je Kapitel und ausführliche Regressionstabellen. Alle Online-Anhänge sind im Text mit dem Suffix *web* referenziert.

Das Skalenhandbuch mit den Beschreibungen aller Erhebungsinstrumente wird voraussichtlich im Jahr 2027 veröffentlicht. Dann werden auch die Daten selbst für Sekundäranalysen zur Verfügung gestellt, wie gewohnt beim Forschungsdatenzentrum (FDZ) des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) in Berlin.

14.1 Einführung in die methodischen Grundlagen

Philipp Sterner

Die seit dem Jahr 2000 durchgeführte PISA-Studie gilt als weltweit größte Studie zur Feststellung der Ergebnisse von Bildungssystemen in den teilnehmenden ▶ Staaten. Sie findet in einem dreijährigen (bzw. ab 2025 in einem vierjährigen) Abstand statt und misst die Leistungsfähigkeit eines Bildungssystems hinsichtlich der Kompetenzen von 15-jährigen Jugendlichen in den drei Kernbereichen Lesen, Mathematik und Naturwissenschaften. Diese ▶ Kerndomänen werden um weitere Kompetenzen aus wechselnden Bereichen ergänzt, die sogenannte ▶ innovative Domäne; speziell um Problemlösen (problem solving) in den PISA-

1 Autor*innen der Kapitel 14.1 bis 14.6

Runden 2003 und 2012 (OECD, 2013), kollaboratives Problemlösen (collaborative problem solving) in der PISA-Runde 2015 (OECD, 2017), globale Kompetenz (global competence) in der PISA-Runde 2018 (OECD, 2019), kreatives Denken (creative thinking) in der PISA-Runde 2022 (OECD, 2023), sowie Lernen in der digitalen Welt (Learning in the Digital World) in der aktuellen Runde 2025. Darüber hinaus gibt es die international optionalen Zusatzdomänen Finanzkompetenz (Financial Literacy) und Fremdsprachenkompetenz Englisch (Foreign Language Assessment).

Mittlerweile nehmen über 90 Staaten an PISA teil, darunter alle OECD-Mitgliedsstaaten, OECD-Partnerstaaten, einzelne Ökonomien sowie alle Mitglieder der Europäischen Union. Das Testalter der teilnehmenden Schüler*innen ist für alle Staaten 15 Jahre, da Schüler*innen in den meisten teilnehmenden Staaten in diesem Alter kurz vor dem Abschluss der Pflichtschulzeit stehen; die exakte Altersspannweite war von 15 Jahren und drei vollendeten Monaten bis 16 Jahre und zwei vollendete Monate zum Startpunkt des Testzeitraums. Durch die Definition der zu testenden ▶ Population über das Testalter anstelle einer spezifischen zu testenden Klassenstufe besteht die deutsche PISA-Stichprobe aus Schüler*innen der Jahrgangsstufen 7 bis 11. Dies ist ein Abgrenzungsmerkmal zu anderen internationalen Bildungsvergleichsstudien, in denen die Zielpopulation durch eine oder mehrere Jahrgangsstufen definiert wird.

Zur Sicherstellung fairer, vergleichbarer und valider Messungen werden bei PISA Methoden verwendet, die hinsichtlich der ▶ Item- und Skalenkonstruktion, der Stichprobenziehung sowie der psychometrischen Skalierungsverfahren höchsten internationalen Standards entsprechen. In den technischen Standards zu PISA sind die verbindlichen OECD-Vorgaben zur standardisierten Durchführung der Erhebung zusammengefasst (OECD, 2026a). Sie beschreiben unter anderem Anforderungen an die Stichprobe, die Übersetzung der Instrumente, die Testdurchführung, die Qualitätskontrolle und die Datenaufbereitung, und geben damit das hochstandardisierte Gesamtvorgehen vor. Sie dienen auch als Maßstab für die Bewertung der Datenqualität (die sog. ▶ Data Adjudication), also für die Entscheidung, ob die Daten eines Teilnehmerstaates methodisch hinreichend belastbar sind, um in die internationalen Berichte aufgenommen zu werden. Hervorzuheben ist, dass in Deutschland alle Vorgaben der PISA-Erhebungsrunde 2025 eingehalten wurden und die vorliegenden Daten damit sowohl im internationalen Vergleich als auch im Zeitverlauf über Erhebungsrunden ohne methodische Einschränkungen interpretiert werden dürfen. In Abgrenzung zu den technischen Standards gibt es außerdem noch den Technical Report (technischen Bericht) der OECD (OECD, 2026b). In diesem sind die konkrete Umsetzung und die methodisch-statistischen Verfahren, zum Beispiel bezüglich Testdesign und Skalierung, einer Erhebungsrunde dokumentiert.

Die Abschnitte in diesem Kapitel orientieren sich grob am Ablauf und Vorgehen der Tätigkeiten des ZIB an der TU München, das mit der ▶ nationalen Studienleitung beauftragt ist. Wichtig ist hierbei zu betonen, dass viele Arbeitsschritte bei PISA eher iterativen Charakter haben, weshalb die Abfolge der Abschnitte nicht streng chronologisch verstanden werden sollte. Darüber hinaus werden hier ausschließlich die methodischen Grundlagen und Abläufe der Haupterhebung beschrieben. Im Rahmen der dreijährigen Erhebungsrunde von PISA 2025 wurde auch ein umfangreicher ▶ Feldtest durchgeführt. Dieser diente als Generalprobe, bei der neue und bestehende Aufgaben empirisch geprüft, Übersetzungseffekte untersucht sowie ▶ Testplattform, Stichprobenziehung und Erhebungsabläufe getestet werden konnten. Die Ergebnisse des Feldtests werden vor allem dazu genutzt, die endgültigen Aufgaben, Testformen und administrativen Verfahren für die Haupterhebung festzulegen; repräsentative Vergleiche zwischen teilnehmenden Staaten werden daraus nicht berichtet. Details zum Feldtest finden sich im Technical Report der OECD (OECD, 2026b).

Als zentrale Indikatoren der Leistungsfähigkeit eines Bildungssystems gelten in PISA die oben genannten Kompetenzen der Schüler*innen. Diese Kompetenzen werden mittels aufwändig erstellter und vorab im Rahmen von Feldtests erprobter, standardisierter Tests gemessen. Vor der Testung wurde das inhaltlich-theoretische Rahmenmodell, speziell für die Hauptdomäne Naturwissenschaften, in einem aufwändigen Prozess überarbeitet und entwickelt. Zusätzlich werden über Fragebögen bildungsrelevante Hintergrundinformationen zu den Jugendlichen, ihren Eltern, den Lehrkräften und den Schulen (erfasst über die Schulleitungen) erhoben. Neben international verpflichtenden Kompetenzdomänen und Fragebögen gibt es auch optionale nationale Ergänzungen in beiden Bereichen, also sowohl bei der Messung der Kompetenzen als auch bei den Hintergrundfragebögen. In Abschnitt 14.2 wird ein Überblick über den Pflichtteil sowie die in Deutschland angewandten optionalen Module gegeben.

Um eine weltweit faire Messung zu gewährleisten, müssen sowohl die Kompetenztests als auch die Hintergrundfragebögen, die ursprünglich in englischer und französischer Sprache zur Verfügung gestellt werden, in die jeweiligen Landessprachen übersetzt werden. Der Übersetzungsprozess wird in den nationalen Studienleitungen der teilnehmenden Staaten durchgeführt und international von cApStAn beaufsichtigt. Der internationale Auftragnehmer cApStAn begleitet die linguistische Qualitätskontrolle sowie die Übersetzungen aller Materialien bereits seit der ersten Erhebungsrunde im Jahr 2000. Die besondere Herausforderung bei der Übersetzung besteht darin, nicht nur sprachlich äquivalente Versionen zu generieren, sondern auch etwaige kulturelle Besonderheiten in den jeweiligen Testsprachen zu berücksichtigen. Fehler oder mangelnde Gründlichkeit in den Übersetzungen könnten dazu führen, dass Items zur Messung der Kompetenzen oder die Items der Hintergrundfragebögen in verschiedenen Staaten unterschiedliche Schwierigkeitsgrade beziehungsweise testtheoretische Eigenschaften aufweisen. Dies würde wiederum die Fairness der Messung sowie die anschließende Vergleichbarkeit zwischen den Staaten einschränken. Dieser Prozess aus Übersetzung und Anpassung wird ebenfalls in Abschnitt 14.2 beschrieben.

Die Zielpopulation in PISA sind 15-jährige Schüler*innen. Aus der Zielpopulation werden in den teilnehmenden Staaten repräsentative Stichproben gezogen. Damit sich Analysen und Aussagen aus der jeweiligen Stichprobe auch auf die gesamte Population übertragen lassen, ist es unabdingbar, dass die Stichprobe nach einem einheitlichen und standardisierten Plan gezogen wird. Der zweistufige Ablauf der Stichprobenziehung wird in Abschnitt 14.3 erklärt. Zudem finden sich dort die vorgegebenen und realisierten Teilnahmequoten.

Neben der Ziehung der Stichprobe ist auch die Testdurchführung in den gezogenen Schulen maßgeblich für die Qualität der erhobenen Daten und der damit verbundenen Aussagen über die Kompetenzen der Schüler*innen. Besonders herausfordernd ist dabei, dass die Testung an allen Schulen in Deutschland sowie international nach gleichen Vorgaben, also hochstandardisiert, durchgeführt werden muss. In Abschnitt 14.4 werden die Organisation und die praktische Umsetzung der Testung durch die IEA in Hamburg in der PISA-Erhebungsrunde 2025 in Deutschland beschrieben.

Bei der Testung selbst werden die Kompetenzitems und Hintergrundfragebögen in einem Testdesign aufbereitet und allen Teilnehmenden dargeboten. In den meisten Staaten, so auch in Deutschland, erfolgt die Durchführung der Kompetenztests bereits seit PISA 2015 vollständig am Computer. Daraus ergeben sich aus testtheoretischer Sicht erhebliche Vorteile, allen voran die Möglichkeit des adaptiven Testens. Dieses Prinzip beruht auf der Auswahl des nächsten zu lösenden Items anhand eines Vergleichs zwischen ▶ Itemschwierigkeit und der zu diesem Zeitpunkt gemessenen Kompetenz einer Person. Sobald die Daten erhoben sind, müssen die Itemantworten der Schüler*innen zu individuellen Kompetenzwerten in den Domänen aggregiert werden. Hierbei ist zu beachten, dass die Items unterschiedliche Schwierigkeitsgrade aufweisen und dass aufgrund der Verwendung ge-

plant-unvollständiger ▶ Testhefte im Testdesign nicht alle Schüler*innen dieselben Items erhalten. Darüber hinaus bearbeiten auch nicht alle Schüler*innen die gesamte Menge an Aufgaben, sondern nur einen Bruchteil davon, um die Gesamttestzeit in einem zumutbaren Rahmen zu halten. Dementsprechend ist das bloße Aufsummieren der richtigen Antworten (wie etwa in einer Klassenarbeit) keine geeignete Möglichkeit, um Kompetenzen zu quantifizieren. Stattdessen wird zur Skalierung der Kompetenzmessungen auf testtheoretische Modelle der ▶ Item-Response-Theorie (IRT) zurückgegriffen, die den speziellen Rahmenbedingungen von ▶ Large-Scale Assessments wie PISA Rechnung tragen. Darüber hinaus werden einige Items der Hintergrundfragebögen zu Indizes (z. B. einer Quantifizierung des sozioökonomischen Hintergrunds) zusammengefasst. Diese Zusammenfassung basiert teilweise auch auf Skalierungsmodellen der IRT. Das genaue Vorgehen bei der Skalierung der Kompetenzmessungen sowie bei der Aggregation von Items aus Hintergrundfragebögen zu einzelnen Indizes wird in Abschnitt 14.5 beschrieben.

Diese Skalierung wurde durch den ▶ internationalen Auftragnehmer ACER vorgenommen. Nach Abschluss dieser Datenaufbereitung stellte ACER dem ZIB als nationaler Studienleitung für Deutschland im Frühjahr 2026 erstmals Daten zur Verfügung. Die Kompetenzmessungen werden für alle Schüler*innen mittels ▶ Plausible Values (plausible Werte) operationalisiert (s. ebenfalls Abschnitt 14.5). Aufgrund der Art und Weise, wie plausible Werte statistisch ermittelt werden, kann es dabei nicht zu fehlenden Werten kommen; alle Schüler*innen erhalten für alle Kompetenzen jeweils zehn plausible Werte. Bei Antworten auf Einzelitems in den Hintergrundfragebögen und daraus resultierenden Indizes kann es allerdings zu fehlenden Werten kommen, zum Beispiel aufgrund des Testdesigns (geplante fehlende Werte) oder weil eine Person ein Item nicht beantworten wollte. Der Umgang mit fehlenden Werten ist statistisch komplex. In Deutschland wurde die Entscheidung getroffen, fehlende Werte auf Ebene der Indizes zu imputieren, also ebenfalls durch plausible Werte basierend auf einem Imputationsmodell zu ersetzen. Dieses Vorgehen sowie die genaueren Hintergründe dieser Entscheidung werden in Abschnitt 14.6 beschrieben.

14.2 Erhebungsinstrumente: Internationale Vorgaben, nationale Ergänzungen und Adaption

Jennifer Diedrich, Elisabeth González Rodríguez & Maren Müller

PISA dient dem internationalen Vergleich von Bildungssystemen. Damit die Ergebnisse zwischen den teilnehmenden Staaten vergleichbar sind, müssen nicht nur die Abläufe der Erhebung nach gemeinsamen Standards erfolgen, etwa die Testdurchführung (siehe Abschnitt 14.4), sondern auch die eingesetzten Instrumente aufeinander abgestimmt sein. Dazu zählen die Kompetenztests ebenso wie die Kontextfragebögen. Die Arbeit der nationalen Studienleitungen beginnt daher in jeder Erhebungsrunde mit der Begutachtung der internationalen Rahmenkonzeptionen und setzt sich mit der Übersetzung und Adaption der daraus entwickelten Aufgaben und Fragebogeninstrumente fort.

Auch in Deutschland folgen die Entwicklung und Vorbereitung der Erhebungsinstrumente den internationalen Vorgaben. Zugleich werden die international vorgesehenen Instrumente um nationale Ergänzungen erweitert, wenn dies für die Beantwortung spezifischer Fragestellungen in Deutschland erforderlich ist. Dazu gehören insbesondere Instrumente für nationale ▶ Trendanalysen, ergänzende Kontextmerkmale sowie Module aus Begleitforschungsprojekten. Die in Deutschland vergleichsweise breite Begleitforschung in PISA spiegelt sich daher auch in der Anzahl und Vielfalt der eingesetzten Instrumente wider.

Ein weiterer zentraler Bestandteil der Instrumentenvorbereitung ist die Übersetzung und Adaption der internationalen Materialien. Die nationale Studienleitung am ZIB arbeitet da-

bei eng mit den nationalen Studienleitungen aus Luxemburg, Österreich und der Schweiz zusammen. Ziel dieser Kooperation ist es, eine deutsche Fassung zu erstellen, die mit den internationalen Vorgaben vereinbar ist und zugleich eine möglichst hohe sprachliche Konsistenz im deutschsprachigen Raum herstellt. Nationale Besonderheiten werden anschließend im Rahmen der Adaption berücksichtigt.

Der folgende Abschnitt beschreibt zunächst die in PISA 2025 in Deutschland eingesetzten Instrumente (Tabelle 14.1). Die Darstellung folgt dabei der Logik von (1) international verpflichtenden Bestandteilen über (2) internationale Optionen hin zu (3) nationalen Ergänzungen, zu denen die beiden Begleitforschungsprojekte PISA innovative Domains in Time 2 (PinDits2) und PISA-Schreiben gehören. Innerhalb dieser Bereiche werden jeweils zunächst die Kompetenztests und anschließend die Fragebögen dargestellt. Zuletzt wird in diesem Abschnitt (4) der Übersetzungs- und Adaptionsprozess beschrieben. Dabei werden zunächst die internationalen Vorgaben skizziert und anschließend die Besonderheiten der deutschsprachigen Übersetzungskooperation erläutert.

Deutschland nahm in PISA 2025 neben den drei Hauptdomänen und der innovativen Domäne Lernen in der digitalen Welt an der internationalen Zusatzdomäne Fremdsprachenkompetenz Englisch teil. Die Ergebnisse dieser Zusatzdomäne werden im Sommer 2027 veröffentlicht. Der vorliegende Abschnitt beschreibt dennoch alle eingesetzten Instrumente, um ein vollständiges Bild der Erhebung in Deutschland zu vermitteln.

Tabelle 14.1: Übersicht der Instrumente in PISA 2025 in Deutschland

Instrumentenbereich	Internationale Pflichtbestandteile (IP)	Internationale Optionen (IO)	Nationale Ergänzungen	
			ZIB-Modul (ZM)	Weitere Ergänzungen (NE)
Kognitive Testinstrumente (K)	IPK1: Naturwissenschaftliche Kompetenz[a]	IOK1: Zusatzdomäne Fremdsprachenkompetenz Englisch[d]	ZMK1: Berliner Test zur Erfassung fluider und kristalliner Intelligenz[e]	
	IPK2: Lesekompetenz		ZMK2: Problemlösekompetenz aus PISA 2003[e]	
	IPK3: Mathematische Kompetenz		ZMK3: Kreatives Denken aus PISA 2022[e]	
	IPK4: Lernen in der digitalen Welt		ZMK4a: Schreibkompetenz Englisch ohne KI[d,f]	
			ZMK4b: Schreibkompetenz Englisch mit KI[d, f]	
Kontextfragebögen (F)	IPF1: Schüler*innen[c]	IOF1: Eltern[b]	ZMF1: Fragebogen zur Schreibkompetenz Englisch	NEF1: Nationale ergänzte Fragen in IPF1 (beginnend immer mit der Nummer ST8**)
	IPF2: Schulleitungen[b]	IOF2: Lehrkräfte[c]		NEF2: Zusatzfragebogen zum fachdidaktischen Wissen der Englischlehrkräfte
		IOF3: IKT-Modul		NEF3: Zusatzfragebogen zum fachdidaktischen Wissen der Naturwissenschaftslehrkräfte

Anmerkungen: [a] Inklusive umweltwissenschaftliche Kompetenz. [b] Beinhaltete auch Fragen zur Fremdsprachenkompetenz Englisch. [c] Es lagen drei Parallelversionen vor: für die Kernstichprobe sowie für die Stichprobe der Zusatzdomäne Fremdsprachenkompetenz Englisch sowie eine gemeinsame Version für Naturwissenschaften und Englisch. [d] Wurde nur der Stichprobe der Zusatzdomäne vorgegeben. [e] Bestandteil des Begleitforschungsprojekts PinDits 2. [f] Bestandteil des Begleitforschungsprojekts PISA-Schreiben.

14.2.1 Internationale Pflichtbestandteile

Das Kernelement der PISA-Studie bilden die Kompetenztests in den drei wiederkehrenden Kerndomänen sowie in der jeweiligen innovativen Domäne. In PISA 2025 wurden die naturwissenschaftliche Kompetenz (IPK1[2]; Kapitel 1; OECD, 2026c), die Lesekompetenz (IPK2; Kapitel 2; OECD, 2019), die mathematische Kompetenz (IPK3; Kapitel 3; OECD, 2023) sowie die innovative Domäne Lernen in der digitalen Welt erhoben (IPK4; Kapitel 4; OECD, 2026c), in der das modellbasierte Problemlösen, welches in diesem Berichtsband berichtet wird, erfasst wurde.

Die naturwissenschaftliche Kompetenz bildete in PISA 2025 die Hauptdomäne und wurde daher vertieft erhoben. Erstmals bestanden die Rahmenkonzeption und die Erhebungsinstrumente in diesem Bereich aus zwei Teilen: naturwissenschaftliche Kompetenz und umweltwissenschaftliche Kompetenz. Beide ▶ Kompetenzbereiche wurden getrennt erhoben und werden in diesem Berichtsband getrennt berichtet (Kapitel 1).

Ergänzend zu den Kompetenztests werden verpflichtende Kontextfragebögen eingesetzt, um die Kontextbedingungen des Lehrens und Lernens zu erfassen OECD, 2026c). Diese Fragebögen richten sich an Schüler*innen (IPF1) und Schulleitungen (IPF2). Die erhobenen ▶ Merkmale decken zentrale Elemente des Context-Input-Process-Output-Models (Kontext-Ressourcen-Prozesse-Ergebnisse-Modell) nach Scheerens (2016) ab. Zu den Kontextmerkmalen gehört beispielsweise die Schulgröße (Schulleitungsfragebogen; Kapitel 10). Inputmerkmale umfassen unter anderem die soziale Herkunft (Schüler*innenfragebogen; Kapitel 5). Prozessmerkmale beziehen sich beispielsweise auf die Häufigkeit der Nutzung digitaler Medien im Unterricht (Schüler*innenfragebogen; Kapitel 11). Ergebnisbezogene Merkmale werden unter anderem über motivationale Konstrukte wie Freude und Interesse an Naturwissenschaften abgebildet (Schüler*innenfragebogen; Kapitel 9).

Insgesamt umfassen die Kontextfragebögen sowohl domänenspezifische Merkmale, die sich auf die jeweilige Haupt- oder innovative Domäne beziehen, als auch domänenübergreifende Personen- und Kontextmerkmale. Beispiele hierfür sind selbstreguliertes Lernen im Kontext der innovativen Domäne Lernen in der digitalen Welt sowie die Lebenszufriedenheit als domänenübergreifendes Merkmal.

Auf Ebene der PISA-Teilnehmerstaaten ist der Einsatz des Schüler*innenfragebogens (IPF1) und des Schulleitungsfragebogens (IPF2) obligatorisch. In Deutschland variiert der Verpflichtungsgrad zur Beantwortung der Fragebögen jedoch zwischen den Bundesländern (Tabelle 14.2) und hängt vom jeweiligen Schulgesetz ab. Verpflichtend bedeutet, dass die jeweiligen Fragebögen von den ausgewählten Befragten vollständig zu bearbeiten sind. Teilverpflichtend bedeutet, dass Angaben zu Schule und Unterricht verpflichtend sind, während die Beantwortung personenbezogener Fragen freiwillig ist. Ist die Teilnahme am Schüler*innenfragebogen freiwillig, geben die Eltern vorab eine Einverständniserklärung ab.

2 Die Abkürzungen der Erhebungsinstrumente setzen sich zusammen aus dem Verpflichtungsgrad (z. B. IP = Internationale Pflichtbestandteile) und dem Typus des Erhebungsinstruments (K= Kognitive Testinstrumente vs. F = Fragebögen) und sind innerhalb der Zellen fortlaufend nummeriert (s. Überschriften in der Tabelle 14.1.)

Tabelle 14.2: Übersicht der Instrumente in PISA 2025 in Deutschland nach Verpflichtungsgrad in den Bundesländern

	Öffentliche Schulen					Schulen in freier Trägerschaft				
Bundesland	Kompetenz-tests (IPK1–4)	Schüle-r*innen-fragebogen (IPF1)	Schul-leitungs-fragebogen (IPF2)	Lehrkräfte fragebogen (IOF1)	Eltern-fragebogen (IOF2)	Kompetenz-tests (IPK1–4)	Schüle-r*innen-fragebogen (IPF1)2	Schul-leitungs-fragebogen (IPF2)3	Lehrkräfte-fragebogen (IOF1)4	Eltern-fragebogen (IOF2)5
Baden-Württemberg	V	V	V	V	F	F	F	F	F	F
Bayern	V	F	F	F	F	V	F	F	F	F
Berlin	V	V	TV	TV	F	F	F	F	F	F
Brandenburg	V	V	V	V	F	–	–	–	–	–
Bremen	V	V	TV	TV	F	–	–	–	–	–
Hamburg	V	V	TV	TV	F	–	–	–	–	–
Hessen	V	V	TV	TV	F	–	–	–	–	–
Mecklenburg-Vorpommern	V	V	V	V	F	–	–	–	–	–
Niedersachsen	V	V	V	V	F	–	–	–	–	–
Nordrhein-Westfalen	V	F	V	V	F	V	F	V	V	F
Rheinland-Pfalz	V	F	TV	TV	F	F	F	F	F	F
Saarland	V	F	TV	TV	F	–	–	–	–	–
Sachsen	V	F	F	F	F	F	F	F	F	F
Sachsen-Anhalt	V	V	V	V	F	–	–	–	–	–
Schleswig-Holstein	V	V	V	V	F	–	–	–	–	–
Thüringen	V	V	V	V	F	–	–	–	–	–

Anmerkung: V = verpflichtend; TV = teilverpflichtend; F = freiwillig; – = nicht vorgesehen bzw. nicht zutreffend.

14.2.2 Internationale Optionen

In PISA 2025 wurde die Fremdsprachenkompetenz Englisch (FLA) erstmals als internationale Option angeboten (IOK1; OECD, 2021). Diese Zusatzdomäne umfasst einen Kompetenztest in den Bereichen Sprechen, Hörverstehen und Lesen sowie ergänzende Module in den Kontextfragebögen. Da sie nicht Teil der regulären PISA-Testung ist, wurde hierfür eine separate Stichprobe gezogen, die in Deutschland knapp 2.000 Schüler*innen umfasste. Im internationalen Design bearbeiteten diese Schüler*innen Aufgaben zur Fremdsprachenkompetenz Englisch (IOK1) sowie zur Lesekompetenz (IPK2). Die Ergebnisse dieser Domäne werden im Mai 2027 veröffentlicht. Während in den Erhebungsrunden von PISA 2012 bis PISA 2022 ausschließlich Finanzkompetenz als internationale Zusatzdomäne angeboten wurde, sollen Finanzkompetenz und Fremdsprachenkompetenz Englisch künftig im Wechsel erhoben werden; in PISA 2029 ist entsprechend wieder Finanzkompetenz vorgesehen. Deutschland hatte in der Vergangenheit nicht an Erhebungen der Finanzkompetenz teilgenommen.

Neben der Zusatzdomäne Fremdsprachenkompetenz Englisch wurden in Deutschland auch mehrere optionale Kontextfragebögen eingesetzt. Dazu gehörten der Elternfragebogen (IOF1), der Lehrkräftefragebogen (IOF2) sowie der Informations- und Kommunikationstech-

nologien Fragebogen (IKT-Modul) für Schüler*innen (IOF3; OECD, 2026c). Auswertungen zu den optionalen Kontextfragebögen finden sich in Teil IV dieses Berichtsbands.

Der Elternfragebogen (IOF1) erfasst seit PISA 2006 das häusliche Lernumfeld der Jugendlichen. In PISA 2025 enthielt der Fragebogen zusätzlich Fragen zum fremdsprachlichen Lernen im häuslichen Umfeld. Im internationalen Design wurden spezifische Fragen zur Fremdsprachenkompetenz Englisch nur den Eltern jener Schüler*innen vorgelegt, die in dieser Zusatzdomäne getestet wurden; in Deutschland erhielten dagegen alle Eltern der getesteten Schüler*innen diese Fragen. Der Elternfragebogen wurde papierbasiert in deutscher Sprache an die Eltern ausgeteilt. Die Beantwortung erfolgte zu Hause, anonym und freiwillig.

Seit der ersten Erhebungsrunde im Jahr 2000 kann der Kontextfragebogen der Schüler*innen um ein Modul zu Informations- und Kommunikationstechnologien (IKT-Fragebogen) erweitert werden, in dem Schüler*innen zu ihren Kenntnissen und ihrer Nutzung digitaler Medien befragt werden (IOF3; OECD, 2026c). Diese Option wurde, wie in vergangenen Erhebungsrunden, auch in Deutschland eingesetzt.

Mit dem Lehrkräftefragebogen (IOF2; internationale Option seit 2015) wurden alle Englischlehrkräfte aus den an PISA teilnehmenden Schulen befragt, damit diese Lehrkräfte mit den Schüler*innen verknüpft werden können (Abschnitt 14.3). Zusätzlich wurden die Lehrkräfte der Hauptdomäne Naturwissenschaften an jeder teilnehmenden Schule befragt (Abschnitt 14.3.1). Die Befragung erfolgte anonym über eine geschützte Onlineplattform. Der Fragebogen erfasste sowohl strukturelle Aspekte, etwa Arbeitszeiten, als auch motivational-emotionale Merkmale, etwa die Freude an naturwissenschaftlichen Themenfeldern der Lehrkräfte.

14.2.3 Nationale Ergänzungen

Über die internationalen Pflichtbestandteile (IPK1 bis IPF2) und Optionen (IOK1 bis IOF3) hinaus wurden in PISA 2025 mehrere nationale Ergänzungen vorgenommen. Sie erweitern das internationale Erhebungsprogramm um für Deutschland relevante Fragestellungen, insbesondere zu Trends, nationalen Kontextmerkmalen und Begleitforschung. Zwei dieser Ergänzungen waren als eigenständige Begleitforschungsprojekte angelegt: PISA innovative Domains in Time 2 (PinDits2) in der regulären Stichprobe und PISA-Schreiben in der Zusatzstichprobe zur Fremdsprachenkompetenz Englisch (IOK1). Ihre Einbettung in die Durchführung der PISA-Studie wird in Tabelle 14.3 schematisch dargestellt. Weitere nationale Ergänzungen betreffen die Fragebögen und werden im Anschluss beschrieben.

Tabelle 14.3: Schematischer Ablauf der Begleitforschungsstudien

Testtag		Instrumenttyp	Kernstichprobe und Modalklasse	Zusatzstichprobe Fremdsprachenkompetenz
Testtag 1	Kognitive Testinstrumente		Kompetenztest (bis zu 2 Domänen aus IPK1 bis IPK4)	Kompetenztest (IOK1)
	Fragebogen		Schüler*innenfragebogen (IPF1)	Schüler*innenfragebogen (IPF1)
Testtag 2	nationale Ergänzung / Begleitforschung	Kognitive Testinstrumente	PinDits2 (2 Module aus ZMK1 bis ZMK3)	PISA-Schreiben (ZMK4a oder ZMK4b)
		Kognitive Testinstrumente	*[entfällt]*	Kompetenztest (1–2 Domänen aus IPK1–3)
		Kognitive Testinstrumente	*[entfällt]*	PISA-Schreiben II (ZMK4a oder ZMK4b)
		Fragebogen	*[entfällt]*	PISA-Schreiben III (ZMF1)

Anmerkungen: Abkürzungen s. Tabelle 14.1; Blautöne markieren Kompetenztests, Beigetöne Fragebögen; dunklere Töne markieren international verpflichtende Instrumente, hellere Töne nationale/optionale Erhebungsinstrumente

Für den Einsatz in PISA mussten die nationalen Ergänzungen nicht nur inhaltlich entwickelt, sondern auch technisch in die internationale Testumgebung eingebunden werden. Die Testinstrumente beider Projekte wurden hierfür im sogenannten ZIB-Modul umgesetzt und über den Itembuilder administriert (Kroehne, 2023). Der Itembuilder ist eine Testanwendung zur Entwicklung und Ausspielung computerbasierter Aufgaben. Damit die nationalen Ergänzungen gemeinsam mit den internationalen Instrumenten (IPK1 bis IOF3) über die offizielle Testplattform eingesetzt werden konnten, wurden die im Itembuilder entwickelten Inhalte anschließend in die TAO-basierte Testplattform des internationalen Auftragnehmers ACER implementiert.

Innovative Domänen wurden in PISA bislang stets nur in einer einzelnen Erhebungsrunde eingesetzt, das heißt, dass sich die innovative Domäne 2012 (komplexes Problemlösen) in Rahmenkonzeption und Aufgaben von jener in PISA 2015 (kollaboratives Problemlösen) unterschied. Mit PinDits2 werden erstmals horizontale und vertikale Zusammenhänge zwischen innovativen Domänen untersucht. Horizontal werden Zusammenhänge zwischen mehreren innovativen Domänen innerhalb derselben PISA-Stichprobe analysiert. Dafür wurden Aufgaben zu den früheren innovativen Domänen Problemlösen (PISA 2003; ZMK2) und Kreatives Denken (PISA 2022; ZMK3) gemeinsam mit einem Test zur allgemeinen kognitiven Grundfähigkeit eingesetzt (Berliner Test zur Erfassung fluider und kristalliner Intelligenz; ZMK1; Wilhelm et al., 2014). Zusätzlich können diese Befunde mit der innovativen Domäne Lernen in der digitalen Welt aus PISA 2025 (IPK4) in Beziehung gesetzt werden. Vertikal wird es für die Domänen Problemlösen (ZMK2) und Kreatives Denken (ZMK3) erstmals möglich, Entwicklungen über mehrere Erhebungsrunden hinweg zu analysieren. Dazu wurden die jeweiligen Testinstrumente aus früheren Erhebungsrunden im Itembuilder nachgebildet.

In der Zusatzdomäne Fremdsprachenkompetenz Englisch (IOK1) wurde am ersten Testtag die Englischkompetenz der Schüler*innen in den Bereichen Lesen, Hörverstehen und Sprechen erfasst. Der vierte Kompetenzbereich, das Schreiben, wurde in Deutschland national ergänzt. Im Begleitforschungsprojekt PISA-Schreiben wird untersucht, auf welche Schritte des Schreibprozesses sich die Bereitstellung von Unterstützungswerkzeugen auf Basis künstlicher Intelligenz auswirkt. Zu diesem Zweck bearbeiteten alle Schüler*innen der Zu-

satzstichprobe an zwei aufeinanderfolgenden Tagen jeweils zwei Schreibaufgaben: einmal ohne Zugang zu einem KI-basierten Schreibwerkzeug (ZMK4a) und einmal mit Zugang zu einem KI-basierten Schreibwerkzeug (ZMK4b). Die Reihenfolge der Bedingungen wurde intraindividuell variiert. Die Module ZMK4a und ZMK4b wurden ZIB-intern im Itembuilder entwickelt und anschließend über die TAO-basierte Testplattform des internationalen Auftragnehmers ACER ausgespielt.

Am zweiten Testtag bearbeiteten die Schüler*innen der Zusatzstichprobe zusätzlich reguläre PISA-Kompetenzaufgaben in Lesen, Mathematik und/oder Naturwissenschaften (IPK1 bis IPK3), die in der Kernstichprobe bereits am ersten Testtag eingesetzt worden waren. Dies war erforderlich, weil das internationale Design für die Zusatzstichprobe nur für einen Teil der Schüler*innen Aufgaben zur Lesekompetenz in Deutsch (IPK3) vorsah. Erst durch diese nationale Ergänzung lassen sich umfassende Zusammenhänge zwischen der Zusatzdomäne Fremdsprachenkompetenz Englisch (IOK1) und den drei PISA-Kerndomänen analysieren.

Neben den testbezogenen Ergänzungen wurden auch mehrere Fragebogenbestandteile national ergänzt. Bestandteil des PISA-Schreiben-Testmoduls am zweiten Testtag war zunächst ein kurzer Fragebogen für die Schüler*innen zur Handhabbarkeit des KI-basierten Schreibwerkzeugs im Schreibprozess (ZMF1). Darüber hinaus wurden im allgemeinen Schüler*innenfragebogen, wie in bisherigen Erhebungsrunden, weitere für Deutschland relevante Merkmale erhoben (NEF1).

Nationale Ergänzungen werden vor allem aus drei zentralen Gründen (s. Kapitel *Die PISA-Studie 2025*) eingesetzt. Erstens wurden einzelne Tests oder Fragebögen aus vergangenen Erhebungsrunden international nicht fortgeführt, die für die Analyse bedeutsamer Verläufe über die Zeit in Deutschland erforderlich sind. Zweitens fehlen im internationalen Instrumentarium einzelne Merkmale, die für nationale Auswertungen relevant sind. Drittens werden bestimmte Merkmale, die für Bildungspolitik und Bildungswissenschaft in Deutschland von besonderem Interesse sind, international nicht erhoben. Vor diesem Hintergrund wurden in PISA 2025 mehrere Instrumente in den Fragebögen national ergänzt. Darüber hinaus wurden im allgemeinen Schüler*innenfragebogen, wie in bisherigen Erhebungsrunden, weitere für Deutschland relevante Merkmale erhoben (NEF1). Die genaue Aufschlüsselung dieser nationalen Ergänzungen wird im Skalenhandbuch dokumentiert, welches voraussichtlich 2027 veröffentlicht werden wird (ZIB, in Vorbereitung). Die nationalen Ergänzungen in den Fragebögen[3] decken verschiedene Bereiche ab. Dazu zählen beispielsweise ergänzende Angaben zum familiären Hintergrund, die zur Bestimmung des Trends in der sozialen Herkunft herangezogen werden (EGP-Klassen; Erikson et al., 1979; Kapitel 5). Auch für Trendanalysen relevante Konstrukte, etwa Freude und Interesse an Mathematik, wurden erneut erhoben. Dieses Konstrukt war bereits in den Erhebungsrunden 2003 und 2012 Bestandteil der Erhebung.

Zusätzlich wurden die Lehrkräfte gebeten, über einen separaten Link einen zweiten Onlinefragebogen auszufüllen. Dieser erfasste fachdidaktisches Wissen und wurde fachspezifisch für Englisch (NEF2; basierend auf FALKO; Krauss et al., 2017) sowie für Physik, Chemie, Biologie oder Mathematik (NEF3; basierend auf PISA-Ceco, Schiepe-Tiska et al., 2026) angeboten. Die Lehrkräfte konnten selbst auswählen, für welches Fach sie den Test bearbeiteten. Der fachdidaktische Wissenstest wurde über einen sicheren Server der IEA Hamburg ausgespielt.

2 Im Skalenhandbuch sind diese daran erkennbar, dass die Fragen eine ID über 800 haben und auf „DEU" enden.

14.2.4 Übersetzungs- und Adaptionsprozess

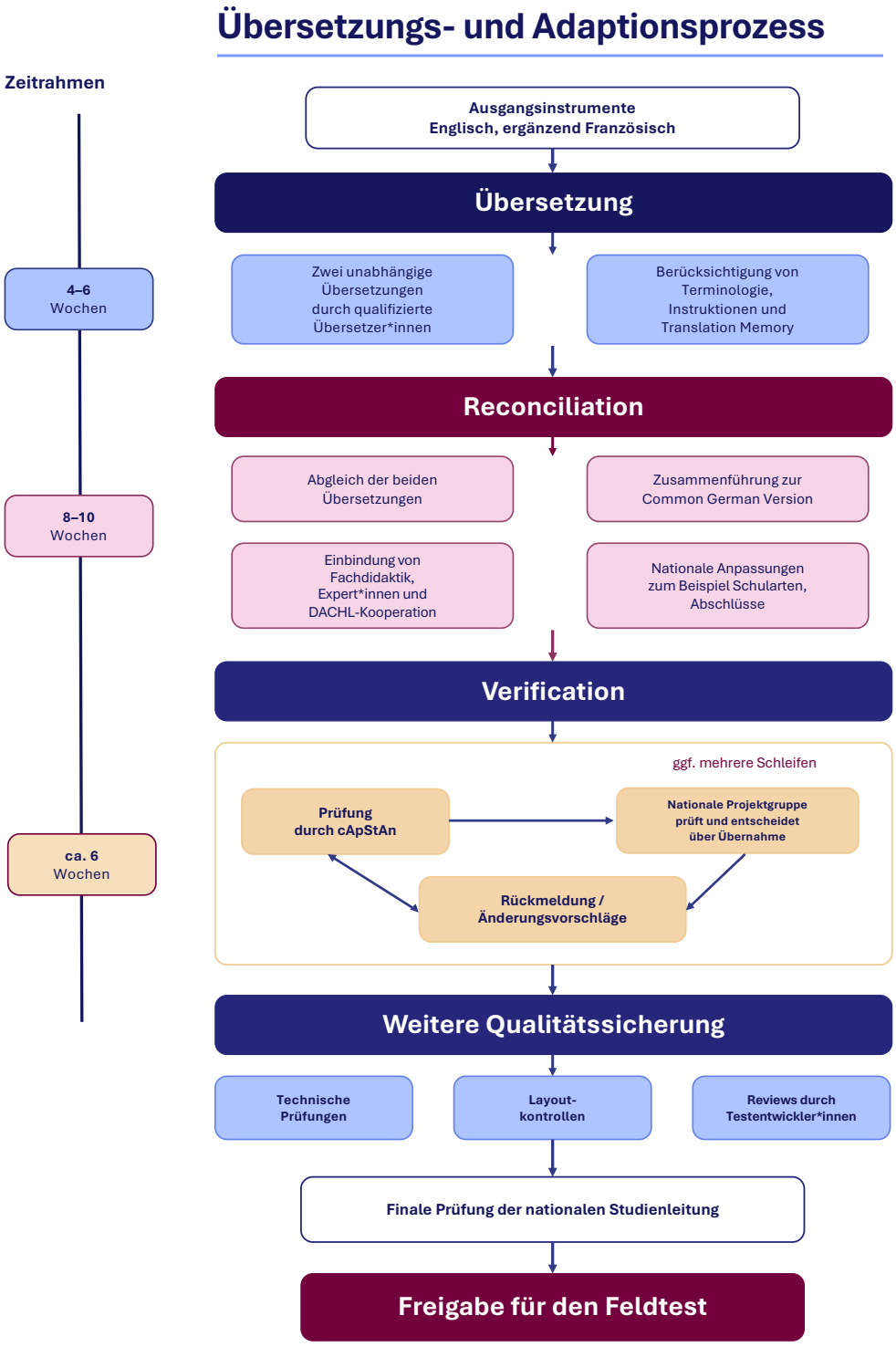
Wie eingangs geschrieben ist die zweite wichtige Aufgabe der nationalen Studienleitung nach der Begutachtung der Rahmenkonzeptionen die Übersetzung beziehungsweise Anpassung an nationale Gegebenheiten. Um die internationale Vergleichbarkeit der Studie zu gewährleisten, sind hier drei Schritte notwendig: (1) die Übersetzung selbst; (2) die Reconciliation (Abgleich) sowie (3) die Verification (Überprüfung).

Abbildung 14.1 fasst den Übersetzungs- und Adaptionsprozess (Anpassung) der PISA-2025-Instrumente zusammen. Ziel dieses mehrstufigen Verfahrens ist es, eine sprachlich präzise, fachlich angemessene und international vergleichbare deutsche Fassung der Instrumente zu erstellen. Startpunkt sind die englischen und ergänzend französischen Instrumente. Diese werden zunächst unabhängig voneinander übersetzt und anschließend im Rahmen der Reconciliation zusammengeführt. Dabei werden fachliche, terminologische und nationale Anforderungen berücksichtigt.

Deutschland kooperiert in diesem Schritt mit den anderen deutschsprachigen Teilnehmerstaaten Luxemburg, Österreich und der Schweiz. In der Kooperation übernimmt jede nationale Studienleitung die Übersetzung eines Teils der Erhebungsinstrumente. Die nationale Studienleitung in Deutschland übernahm in PISA 2025 die Übersetzung der Aufgaben zur naturwissenschaftlichen Kompetenz (IPK1), der Aufgaben zu Lernen in der digitalen Welt (IPK4), des Lehrkräftefragebogens sowie der Fragebogenskalen zur Fremdsprachenkompetenz Englisch (FLA). Anschließend verständigten sich die beteiligten Studienleitungen im Rahmen der Reconciliation auf eine gemeinsame deutsche Fassung, die sogenannte *Common German Version*. Diese bildet die Grundlage für die weiteren nationalen Adaptionen.

Spezifische nationale Besonderheiten, die nur innerhalb Deutschlands relevant sind, werden anschließend im Rahmen des Adaptionsprozesses berücksichtigt. Dies betrifft beispielsweise Begriffe, die in den deutschsprachigen Staaten unterschiedlich gebräuchlich sind. Solche Adaptionen werden bei den internationalen Auftragnehmern beantragt und dokumentiert. Zum Beispiel ist in Österreich Burschen für Jungen gebräuchlicher, sodass die österreichische Studienleitung an allen Stellen, wo in der Common German Version Jungen steht, eine nationale Adaption bei den internationalen Auftragnehmern beantragt. Die anschließende Verification (Überprüfung) durch den internationalen Auftragnehmer cApStAn ist als iterativer Prozess konzipiert. Rückmeldungen und Änderungsvorschläge werden von der nationalen Studienleitung geprüft und gegebenenfalls in die Instrumente eingearbeitet. Ergänzende technische Prüfungen, Layoutkontrollen und Reviews durch Testentwickler*innen dienen der weiteren Qualitätssicherung, bevor die finale Prüfung bei der nationalen Studienleitung und die Freigabe für den Feldtest der PISA-Studie erfolgen.

Abbildung 14.1: Prozessschritte der Übersetzung und Adaption



Übersetzt werden in jeder Erhebungsrunde nur neue Erhebungsinstrumente. Dazu zählen die neuen Aufgaben der Hauptdomäne Naturwissenschaften (IPK1), alle Aufgaben der innovativen Domäne Lernen in der digitalen Welt (IPK4) sowie neue Skalen in den Fragebögen (IPF1, IPF2 und IOF1 bis IOF3). In der Zusatzdomäne Fremdsprachenkompetenz Englisch (IOK1) mussten nur die Instruktionen, aber nicht die Aufgaben übersetzt werden. Trenditems, die über mehrere Erhebungsrunden hinweg möglichst unverändert bleiben müssen, dürfen nur in begründeten Einzelfällen angepasst werden, etwa wenn eindeutige Fehler in vorherigen Übersetzungen erst jetzt auffallen und korrigiert werden müssen, oder wenn sich Teile des allgemeinen Sprachverständnisses ändern (siehe Abschnitt 14.5). Beispielsweise wurde in der Erhebungsrunde 2025 eine in allen PISA-Aufgaben verwendete, fiktive Währung von „Zed" zu „Red" umbenannt, da „Z" seit der großangelegten russischen Aggression gegen die Ukraine als Propagandazeichen konnotiert wird. Die dadurch möglicherweise ausgelösten Gefühle könnten die Messung verzerren.

Im Anschluss an den Feldtest (siehe Abschnitt 14.1) stellt der internationale Auftragnehmer ACER Gütekriterien basierend auf der internationalen und nationalen Stichprobe für alle Aufgaben und Fragebogenskalen zur Verfügung. Wenn sich beispielsweise zeigt, dass in einem Staat eine bestimmte falsche Antwort überdurchschnittlich häufig gewählt wurde, wird die Übersetzung der betreffenden Aufgabe erneut geprüft und gegebenenfalls vor der Haupterhebung angepasst.

14.3 Stichproben und Stichprobenbeschreibung

Claudia Bovensiepen, Lisa Teufele, Sabrina Wagner & Sabine Meinck

Die internationale Vergleichbarkeit und Aussagekraft der PISA-Ergebnisse setzen eine sorgfältige Definition der Zielpopulationen sowie die Ziehung geeigneter Stichproben voraus. Entsprechend bildet die Qualität der Stichprobe eine wesentliche Voraussetzung für belastbare Ergebnisse. Im vorliegenden Abschnitt werden die Zielpopulationen und das Stichprobendesign für Deutschland vorgestellt. Weiterhin werden die Ziehung der Schul-, Schüler*innen- und Lehrkräftestichproben sowie die Teilnahmequoten beschrieben. Abschließend werden die Bildung der Schätzgewichte sowie die Präzision der gewonnenen Daten behandelt.

14.3.1 Populationsdefinitionen und Stichprobendesign

Für die PISA-Studie wurden verschiedene Erhebungen durchgeführt, für die spezifische Zielpopulationen definiert wurden. Im Folgenden werden die Zielpopulationen, das zugrunde liegende Stichprobendesign sowie die daraus abgeleiteten Stichproben beschrieben.

Zielpopulationen

Zur Sicherstellung der internationalen Vergleichbarkeit stammen alle an PISA 2025 teilnehmenden Schüler*innen aus einer international einheitlich definierten Zielpopulation. Diese umfasst 15-jährige Schüler*innen einer teilnehmenden Einheit (z. B. Staat oder Region), die eine Bildungseinrichtung ab der 7. Jahrgangsstufe besuchen. Die konkrete Definition der PISA-Zielpopulation in Deutschland, die im Folgenden als PISA-Kernpopulation bezeichnet wird, besteht aus allen Schüler*innen in den Jahrgangsstufen 7 und höher, die zwischen dem 1. Januar 2009 und dem 31. Dezember 2009 geboren wurden. Somit waren die Schüler*innen der PISA-Kernpopulation zum Testzeitpunkt zwischen 15 Jahren und drei Monaten

sowie 16 Jahren und zwei Monaten alt (OECD, 2026a). Dies ergibt sich aus dem von der ► internationalen Studienleitung festgelegten Testzeitraum.

Ergänzend zur PISA-Kernpopulation wurde in Deutschland die international optionale Zusatzdomäne Fremdsprachenkompetenz Englisch (Foreign Language Assessment English; FLA) zur Erfassung der Englischkompetenz umgesetzt. Englisch gilt dabei als Fremdsprache, sofern es im Bildungssystem formal unterrichtet wird und nicht die primäre Unterrichtssprache ist. Diese Population bildete eine Teilpopulation der PISA-Kernpopulation und umfasste diejenigen Schüler*innen, die vor dem Erhebungsjahr mindestens ein Schuljahr lang Englisch als Fremdsprache im nationalen Bildungssystem gelernt hatten (OECD, 2026a). Darüber hinaus wurde die von der internationalen Studienleitung angebotene nationale Option zur Testung der Modalstufe genutzt. Als PISA-Modalstufe gilt diejenige Jahrgangsstufe, die vom größten Anteil der PISA-Schüler*innen besucht wird; in Deutschland war dies die 9. Jahrgangsstufe. Daher wurde zusätzlich die Population der Neuntklässler*innen definiert und daraus eine neunte Klasse ausgewählt. Diese Stichprobe wird unabhängig vom Alter der Schüler*innen gebildet und ermöglicht zusätzliche Analysen auf Klassenebene.

Die Zielpopulation der naturwissenschaftlichen Lehrkräftebefragung umfasste Lehrkräfte, die im Erhebungsjahr ein naturwissenschaftliches Fach in der sogenannten PISA-Modalstufe unterrichteten. Zusätzlich gab es die Möglichkeit, die Zielpopulation auf Lehrkräfte einer weiteren Jahrgangsstufe auszuweiten, sofern diese ebenfalls mindestens 30 Prozent der PISA-Zielpopulation umfasste (OECD, 2026a). Bei der Umsetzung der internationalen FLA-Option galt eine analoge Definition für die Zielpopulation der Englischlehrkräfte. Diese umfasste alle Lehrkräfte, die im Erhebungsjahr Englisch in der PISA-Modalstufe unterrichteten. Auch hier konnte die Zielpopulation unter den genannten Bedingungen auf eine weitere Jahrgangsstufe ausgeweitet werden (OECD, 2026a).

Schulleitungen der teilnehmenden Schulen sowie Eltern der teilnehmenden Schüler*innen wurden ebenfalls befragt. Diese Befragungspopulationen ergaben sich unmittelbar aus der Schul- beziehungsweise Schüler*innenstichprobe und werden daher im Folgenden nicht gesondert beschrieben.

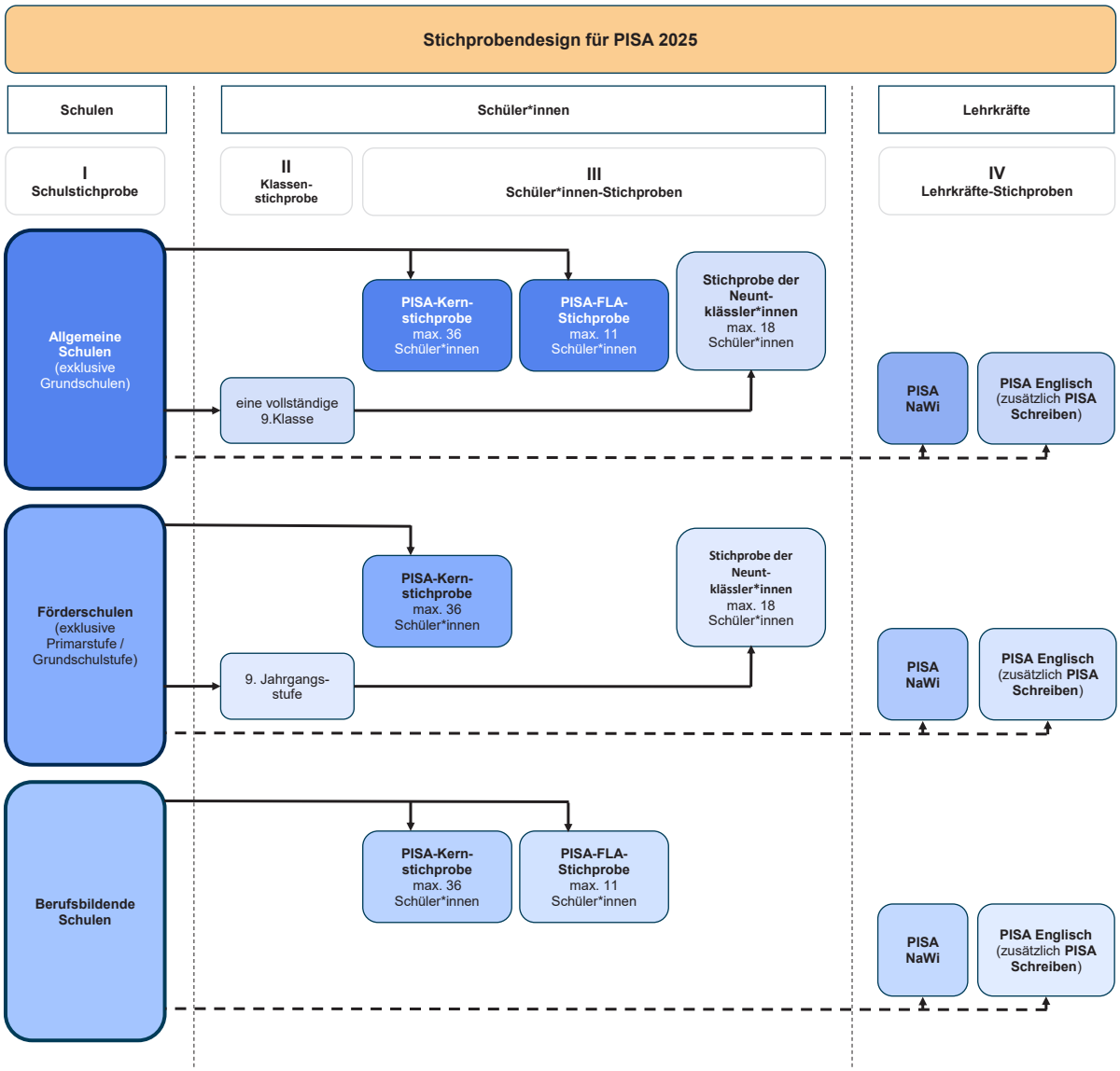
Stichprobendesign

Die Stichprobenziehung im Rahmen der PISA-Studie basiert auf einem zweistufigen, ► stratifizierten Klumpenstichprobendesign, wie es für Studien des Bildungsmonitorings international etabliert ist (Meinck & Vandenplas, 2021; Rutkowski et al., 2014). In der ersten Stufe wurden Schulen ausgewählt, in der zweiten Stufe erfolgte die Ziehung der Schüler*innen innerhalb der ausgewählten Schulen. Bei der praktischen Umsetzung lassen sich jedoch mehrere Ziehungsschritte unterscheiden, die in Abbildung 14.2 schematisch dargestellt sind.

In der ersten Stufe der Stichprobenziehung (Ziehungsschritt I) erfolgte die Auswahl der Schulen als primäre Stichprobeneinheiten aus dem nationalen Schulverzeichnis in Deutschland. Die Durchführung wird im Abschnitt zur *Ziehung der Schulstichprobe* detailliert beschrieben. Die zweite Stufe der Stichprobenziehung umfasste die Ziehungsschritte II bis IV an den ausgewählten Schulen. In Ziehungsschritt II wurde zunächst eine 9. Klasse (Modalstufe) gezogen. In Ziehungsschritt III wurden dann die Untersuchungsstichproben auf Schüler*innenebene festgelegt. Die Auswahl erfolgte auf Grundlage vollständiger schulinterner Listen aller zu den Zielpopulationen gehörenden Schüler*innen beziehungsweise Klassen der 9. Jahrgangsstufe. Die Bestimmung der PISA-Kernstichprobe, der PISA-FLA-Stichprobe sowie der Stichprobe der Neuntklässler*innen erfolgte daher im Rahmen einer integrierten Ziehungsprozedur in festgelegter Reihenfolge: Zunächst wurde die PISA-Kernstichprobe bestimmt, anschließend die PISA-FLA-Stichprobe und danach die Stichprobe der Neuntklässler*innen festgelegt. Bereits gezogene Schüler*innen waren für die Ziehungen der weiteren

Stichproben nicht mehr verfügbar. Durch die festgelegte Ziehungsreihenfolge und dem Ausschluss von Mehrfachziehungen lagen für alle Schüler*innen eindeutig definierte Auswahlwahrscheinlichkeiten vor.

Abbildung 14.2: Schematische Darstellung des Stichprobendesigns von PISA 2025 in Deutschland nach ▶ Schularten mit den Schüler*innenstichproben (PISA-Kernstichprobe, PISA-FLA-Stichprobe und der Stichprobe der Neuntklässler*innen) sowie den angebundenen Lehrkräftebefragungen



Die Vorgaben zur Stichprobengröße für PISA 2025 wurden international festgelegt. Für das in Deutschland verwendete computerbasierte Erhebungsdesign (Computer-based Assessment, CBA) einschließlich FLA war eine Mindeststichprobengröße von insgesamt 7.950 getesteten Schüler*innen vorgesehen, von denen 1.650 Schüler*innen das FLA-Modul bearbeiten sollten (OECD, 2026a). Um diese Zielgröße trotz erwarteter Ausfälle zu erreichen, wurde bei der Stichprobenplanung eine Bruttostichprobe von 10.745 Schüler*innen angesetzt, davon 2.425 für das FLA-Modul. Auf dieser Grundlage wurden die Anzahl der zu ziehenden Schulen sowie die Target Cluster Size (TCS) bestimmt, die festlegt, wie viele Schüler*innen pro teilnehmender Schule in die Schüler*innenstichprobe aufgenommen werden sollten. Für die nationale Option der Neuntklässler*innen bestanden keine internatio-

nal vorgegebenen Mindestanforderungen an die Stichprobengröße. Für diese Teilstichprobe wurde eine Stichprobengröße von 4.165 Schüler*innen vor Ausfällen angestrebt.

In Ziehungsschritt IV wurden die Stichproben der naturwissenschaftlichen Lehrkräfte, der Englischlehrkräfte sowie der Lehrkräfte der Begleitstudie PISA-Schreiben bestimmt. Sie bildeten eigenständige Stichproben, wurden jedoch ausschließlich aus den gezogenen Schulen ausgewählt und nicht über eine separate Schulstichprobe rekrutiert. Detaillierte Beschreibungen der einzelnen Lehrkräftestichproben finden sich im Abschnitt *Stichproben der Lehrkräfte*.

Stichprobe der 15-Jährigen

Für die Stichprobe der PISA-Kernpopulation wurden innerhalb aller für PISA 2025 gezogenen Schulen aus den vollständigen Listen der zur PISA-Kernpopulation gehörenden Schüler*innen (s. Abschnitt *Zielpopulationen*) maximal 36 Schüler*innen zufällig ausgewählt (Abbildung 14.2, Ziehungsschritt III). Die Ziehung der PISA-FLA-Stichprobe erfolgte an allgemeinen und berufsbildenden Schulen, jedoch nicht an Förderschulen (Abbildung 14.2, Ziehungsschritt III). Es wurden höchstens 11 Schüler*innen pro Schule zufällig für diese Stichprobe ausgewählt.

Die tatsächliche Anzahl der gezogenen Schüler*innen für die PISA-Kern- oder PISA-FLA-Stichprobe konnte geringer ausfallen. Die TCS betrug 45 Schüler*innen pro allgemeiner oder berufsbildender Schule. Zwar konnten je Schule bis zu 36 Schüler*innen für die PISA-Kernstichprobe und bis zu 11 Schüler*innen für die PISA-FLA-Stichprobe ausgewählt werden, aufgrund der TCS konnten die maximalen Stichprobengrößen beider Stichproben jedoch innerhalb einer Schule nicht gleichzeitig ausgeschöpft werden. Zudem waren auch Schulen mit weniger als 45 teilnahmeberechtigten Schüler*innen Teil der Stichproben.

Stichprobe der Neuntklässler*innen

Für weiterführende Auswertungen wurden in Deutschland Schüler*innen der 9. Jahrgangsstufe als zusätzliche Population untersucht. Da sich diese Jahrgangsstufe aus Schüler*innen unterschiedlicher Altersgruppen zusammensetzt, umfasste diese Population sowohl 15-Jährige der PISA-Kernpopulation als auch ältere und jüngere Schüler*innen. Dementsprechend konnten Schüler*innen der 9. Klasse je nach Alter entweder der PISA-Kernpopulation, der nationalen Option der Neuntklässler*innen oder beiden Populationen zugeordnet sein. Die sich hieraus ergebenden Überschneidungen bestanden jedoch nur auf Populationsebene. Durch die im Abschnitt *Stichprobendesign* beschriebene integrierte Ziehungsprozedur wurden Mehrfachziehungen derselben Schüler*innen ausgeschlossen.

Die nationale Option der Neuntklässler*innen wurde an allgemeinen Schulen und Förderschulen eingesetzt. An berufsbildenden Schulen erfolgte keine Ziehung, da diese keine 9. Jahrgangsstufe führen (Abbildung 14.2, Ziehungsschritt III). Bei der Listung der Schüler*innen durch die Schule konnten diese durch Angabe der Jahrgangsstufe und des Klassennamens eindeutig einer Klasse zugeordnet werden. Da Förderschulen häufig nur über eine geringe Zahl an Neuntklässler*innen verfügen und hier vielfach keine regulären Klassenstrukturen bestehen, wurden sämtliche Schüler*innen der Jahrgangsstufe 9 einer Schule zu einer virtuellen 9. Klasse zusammengefasst.

Da diese nationale Option kein Bestandteil der internationalen Pflichtbestandteile war, gab es für diese Stichprobe keine international festgelegte TCS. Die Größe der Untersuchungseinheit 9. Klasse pro Schule wurde auf maximal 18 Schüler*innen begrenzt und hing von der Anzahl der in der gezogenen 9. Klasse verbliebenen Schüler*innen ab, die nicht für die PISA-Kern- oder PISA-FLA-Stichproben ausgewählt wurden.

Stichproben der Lehrkräfte

Für die Stichprobe der naturwissenschaftlich unterrichtenden Lehrkräfte (siehe Abbildung 14.2, Ziehungsschritt IV: PISA NaWi) an allgemeinen Schulen und Förderschulen wurden an den teilnehmenden Schulen alle Lehrkräfte einschließlich Referendar*innen gelistet, die im ersten Schulhalbjahr 2024/2025 in der 9. und/oder 10. Jahrgangsstufe die Fächer Biologie, Chemie, Physik und/oder integrierte Naturwissenschaften unterrichteten. An berufsbildenden Schulen wurden Lehrkräfte dieser Fachrichtungen gelistet, die Schüler*innen im ersten Ausbildungsjahr unterrichteten. Alle naturwissenschaftlichen Lehrkräfte an den teilnehmenden Schulen, auf die diese Definition zutraf, wurden gebeten, an der PISA-Befragung teilzunehmen; eine gesonderte Ziehung wurde dementsprechend nicht vorgenommen.

Für die Befragung der Englischlehrkräfte (siehe Abbildung 14.2, Ziehungsschritt IV: PISA-Englisch) listeten die allgemeinen Schulen und Förderschulen alle Lehrkräfte (inkl. Referendar*innen), die im ersten Schulhalbjahr 2024/2025 in den Jahrgangsstufen 9 und/oder 10 Englisch unterrichteten beziehungsweise an berufsbildenden Schulen Schüler*innen des ersten Ausbildungsjahres. Eine gesonderte Ziehung war auch hier nicht vorgesehen, da sämtliche erfassten Englischlehrkräfte in die PISA-Erhebung einbezogen wurden.

Darüber hinaus wurden im Rahmen der Begleitstudie PISA-Schreiben Lehrkräfte befragt (siehe Abbildung 14.2, Ziehungsschritt IV: PISA-Schreiben), die im ersten Schulhalbjahr 2024/2025 an allgemeinen Schulen und Förderschulen Englisch ab der Jahrgangsstufe 5 unterrichteten. An berufsbildenden Schulen wurden Englischlehrkräfte gelistet, die Schüler*innen im zweiten und dritten Ausbildungsjahr unterrichteten. Diese Stichprobe umfasste ausschließlich Englischlehrkräfte, die nicht für die PISA-Lehrkräftebefragung vorgesehen waren. Eine separate Ziehung dieser Lehrkräfte war nicht erforderlich, da alle hierfür gelisteten Lehrkräfte bereits an der nationalen Lehrkräftebefragung teilnahmen.

Von der Listung ausgeschlossen waren Lehrkräfte, die aufgrund ihrer Funktion oder ihres Beschäftigungsstatus nicht zu den definierten Zielpopulationen gehörten. Dies betraf insbesondere Mitglieder der Schulleitung, vorübergehend eingesetzte oder langzeitabwesende Lehrkräfte sowie Lehrbeauftragte, Honorarkräfte und unterrichtsunterstützendes Personal.

Zuordnung von Lehrkräften zu Klassen und Schüler*innen

Für die Lehrkräfteerhebungen wurden unterschiedliche Verfahren zur Verknüpfung der Lehrkräfte-, Klassen- und Schüler*innendaten eingesetzt.

Für die Erhebung der naturwissenschaftlichen Lehrkräfte an allgemeinen Schulen kennzeichneten die Schulen diejenigen Lehrkräfte, die im ersten Schulhalbjahr 2024/2025 in der gezogenen 9. Klasse ein naturwissenschaftliches Fach unterrichtet hatten. Dadurch konnten deren Angaben mit den Daten der Schüler*innen dieser Klasse verknüpft werden. Für die naturwissenschaftlichen Lehrkräfte an berufsbildenden Schulen und Förderschulen wurden keine entsprechenden Zuordnungsinformationen erhoben.

Im Rahmen der FLA-Erhebung ordneten die Schulen an allgemeinen und berufsbildenden Schulen den Schüler*innen der PISA-FLA-Stichprobe die Englischlehrkräfte zu, die sie im ersten Schulhalbjahr 2024/2025 unterrichtet hatten. Bei Team-Teaching konnten den Schüler*innen mehrere Lehrkräfte zugeordnet werden. Die erhobenen Zuordnungsinformationen ermöglichten die Verknüpfung der Daten der FLA-Schüler*innen mit den Angaben der befragten Englischlehrkräfte. An Förderschulen wurden keine entsprechenden Zuordnungsinformationen erhoben.

14.3.2 Implementierung und Realisierung des Stichprobendesigns

Nach der Vorstellung der Zielpopulationen sowie des zweistufigen Stichprobendesigns werden im Folgenden dessen praktische Umsetzung und die daraus resultierenden Stichproben beschrieben. Zunächst werden die Ziehung der Schulen durch den internationalen Auftragnehmer und die realisierte Schulstichprobe behandelt. Anschließend werden die Ziehungen der Schüler*innen sowie die realisierten Schüler*innenstichproben beschrieben. Abschließend werden die Auswahl der Lehrkräftestichproben und deren Realisierung dargestellt.

Ziehung der Schulstichprobe

Als Grundlage für die Ziehung der Schulstichprobe diente eine vollständige Liste der Schulen, die im Jahr 2025 potenziell 15-jährige Schüler*innen unterrichteten. Diese wurde auf Grundlage von Schullisten erstellt, die von den statistischen Landesämtern beziehungsweise den Kultusministerien der Bundesländer bereitgestellt wurden und auf den amtlichen Schulstatistiken des Schuljahres 2022/2023 basierten. Diese Statistiken werden in den meisten Bundesländern jährlich anhand stichtagsbezogener Meldungen aller Schulen erhoben.

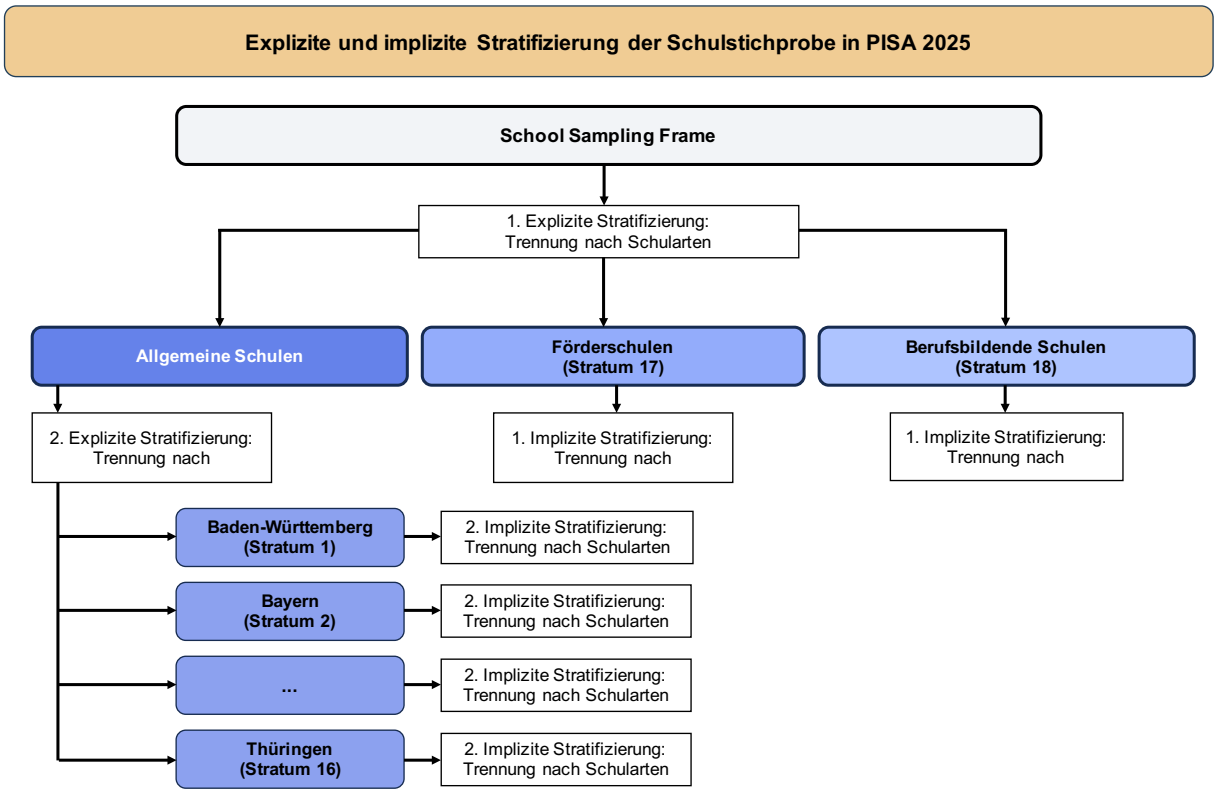
Insgesamt wurden 15.481 Schulen in den Sampling Frame aufgenommen, die Jahrgangsstufe 7 und höher führten, also alle allgemeinen Schulen (exklusive Grundschulen), Förderschulen (exklusive Primar- bzw. Grundschulstufe) und berufsbildende Schulen.

Um eine systematische Auswahl der Schulen zu ermöglichen, wurde im Sampling Frame jeder Schule ein Measure of Size (MOS) hinzugefügt, also eine geschätzte Zahl der 15-Jährigen. Für diese PISA-Erhebungsrunde basierte das MOS auf der Zahl der Schüler*innen mit Geburtsjahrgang 2007, da die amtlichen Schulstatistiken bereits zwei Jahre alt waren und dieser Jahrgang herangezogen werden musste, um die PISA-Kernpopulation mit Geburtsjahrgang 2009 abzuschätzen. Zusätzlich wurde für die Stichprobe der nationalen Option eine Schätzung der Anzahl der Neuntklässler*innen pro Schule in den Sampling Frame aufgenommen.

Zur Vorbereitung der Stichprobenziehung erfolgte eine Gruppierung der Schulen im Sampling Frame nach festgelegten Kriterien, die durch die Stratifizierungsvariablen definiert wurden. Ein Stratum bezeichnet eine Gruppe von Schulen, die durch die Stratifizierung definierten Eigenschaften teilen und sich dadurch idealerweise auch in weiteren Merkmalen wie der durchschnittlichen Kompetenz der Schüler*innen ähneln. Eine Stratifizierung kann explizit oder implizit erfolgen und dient der Sicherstellung der Repräsentativität der Stichprobe in Bezug auf die Population. Für PISA 2025 wurden in Deutschland zwei explizite und zwei implizite Stratifizierungsvariablen eingesetzt (Abbildung 14.3).

Bei einer expliziten Stratifizierung werden Schulen in klar abgegrenzte Strata eingeteilt. Jedes Stratum wird unabhängig von den anderen behandelt und in jedem Stratum erfolgt eine eigene Zufallsziehung. In der ersten expliziten Stratifizierung wurden alle Schulen des Sampling Frames in drei Gruppen eingeteilt: allgemeine Schulen, Förderschulen und berufsbildende Schulen. Anschließend erfolgte die zweite explizite Stratifizierung, bei der die allgemeinen Schulen auf Ebene der Bundesländer unterteilt wurden. Insgesamt ergaben sich 18 explizite Strata: Stratum 1 bis 16 umfassten die allgemeinen Schulen in den 16 Bundesländern, Stratum 17 alle Förderschulen und Stratum 18 alle berufsbildenden Schulen. Die besondere Handhabung der Förder- und beruflichen Schulen wurde aufgrund ihrer unterschiedlichen Verteilung über die Bundesländer und ihrer geringeren Fallzahlen festgelegt.

Abbildung 14.3: Schematische Darstellung des Stratifizierungsdesigns von PISA 2025 in Deutschland.



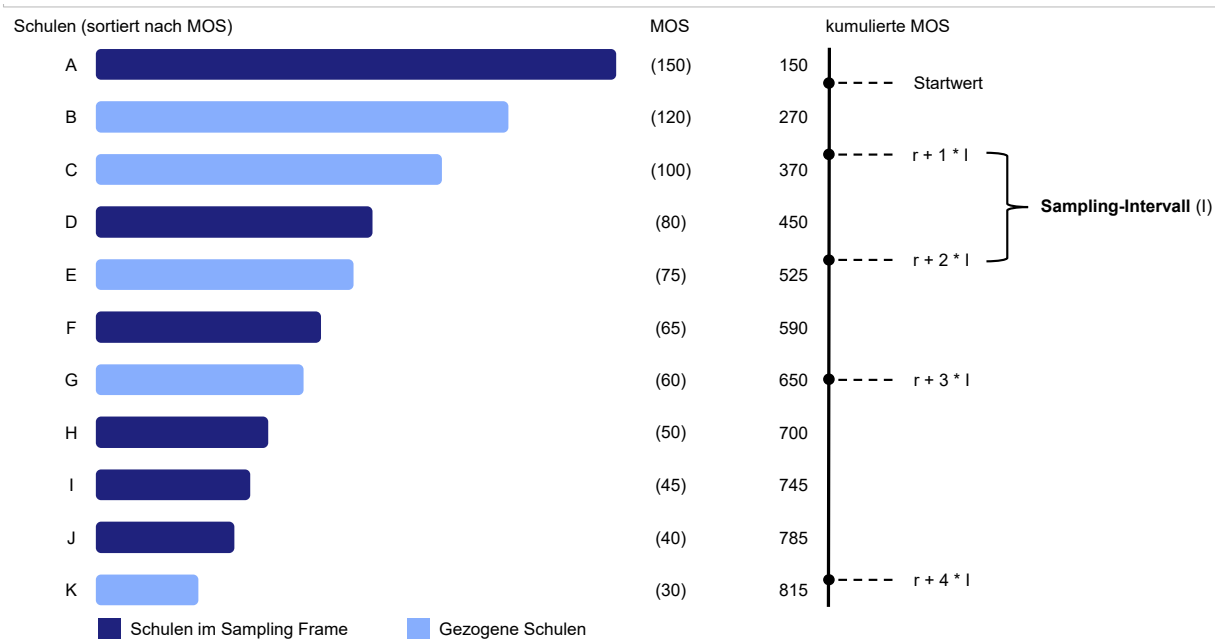
Eine implizite Stratifizierung erfolgt innerhalb eines expliziten Stratums. Dabei werden die Schulen nach implizitem Stratum sowie nach Schulgröße (definiert durch das MOS) sortiert. In Verbindung mit einer systematischen Zufallsziehung führt diese Methode dazu, dass die gezogenen Schulen die verschiedenen Untergruppen innerhalb eines expliziten Stratums ungefähr im gleichen Verhältnis repräsentieren, wie sie im Sampling Frame vorkommen. Förderschulen und berufsbildende Schulen wurden in der ersten impliziten Stratifizierung nach Bundesländern geordnet. Die allgemeinen Schulen in den einzelnen Bundesländern wurden in der zweiten impliziten Stratifizierung nach den Schularten Hauptschule (HS), Realschule (RS), Schule mit mehreren Bildungsgängen (MB), Gymnasium (GY), Integrierte Gesamtschule (IG) und Schule nicht-deutscher Herkunftssprache (ND) unterteilt. ND bezeichnet Schüler*innen mit nicht-deutscher Herkunftssprache, die in Hessen außerhalb von Regelklassen unterrichtet werden. Diese Gruppe wurde in der PISA-Stichprobe als eigene Schulart geführt, um ihren Anteil proportional abzubilden, obwohl sie keine Schulart des deutschen Schulsystems ist.

Zusammenfassend wurde der School-Sampling-Frame der PISA-Studie 2025 mit folgenden Angaben pro Schule erstellt:

- Pseudonymisierte offizielle Schulnummer (nationale Schul-ID)
- Schätzung der Anzahl der Schüler*innen in der PISA-Kernpopulation
- Schätzung der Anzahl der Schüler*innen in einer 9. Klasse
- Erste explizite Stratifizierungsvariable zur Kennzeichnung von allgemeinen Schulen, Förderschulen und berufsbildenden Schulen
- Zweite explizite Stratifizierungsvariable zur Kennzeichnung der Bundesländer bei allgemeinen Schulen
- Erste implizite Stratifizierungsvariable zur Kennzeichnung der Bundesländer bei Förderschulen und berufsbildenden Schulen
- Zweite implizite Stratifizierungsvariable zur Kennzeichnung der Schularten bei allgemeinen Schulen

Auf Grundlage der gebildeten Strata erfolgte die Ziehung der Schulen nach dem PPS-Verfahren (Probability Proportional to Size). Zur Umsetzung dieses Verfahrens wurde innerhalb jedes expliziten Stratums nach Anwendung der impliziten Stratifizierung der Sampling Frame nach der MOS geordnet. Für die Auswahl der Schulen wurde die MOS kumuliert, indem die MOS-Werte der geordneten Schulen fortlaufend aufsummiert wurden. Anschließend wurde ein zufälliger Startwert bestimmt und das Sampling-Intervall berechnet. Dieses ergab sich aus dem Verhältnis der Gesamt-MOS des Stratums zur Anzahl der zu ziehenden Schulen. Ausgehend vom Startwert wurden durch fortlaufende Addition des Sampling-Intervalls Auswahlwerte gebildet. Eine Schule wurde jeweils dann gezogen, wenn ihre kumulierte MOS den entsprechenden Auswahlwert erreichte oder überschritt. Dieses Verfahren ist in Abbildung 14.4 schematisch dargestellt.

Abbildung 14.4: Schematische Darstellung der systematischen Stichprobenziehung mit Wahrscheinlichkeiten, die proportional zur Größe sind (PPS).



Anmerkung: Die Auswahl erfolgt auf Basis der kumulierten MOS unter Verwendung eines zufälligen Startwerts *r* und eines Sampling-Intervalls *I*, aus denen sich die Auswahlpunkte ergeben. Die hellblau hervorgehobenen Balken kennzeichnen die gezogenen Schulen. (adaptiert nach Zuehlke, 2011)

Das PPS-Verfahren führt dazu, dass die Ziehungswahrscheinlichkeiten der Schulen von ihrer MOS abhängen. Entsprechend weisen größere Schulen höhere Ziehungswahrscheinlichkeiten auf als kleinere Schulen. Umgekehrt haben einzelne Schüler*innen an großen Schulen eine geringere Wahrscheinlichkeit, gezogen zu werden, wodurch die höhere Ziehungswahrscheinlichkeit großer Schulen ausgeglichen wird. Durch die Anwendung von PPS wird ein sogenanntes self-weighting Design erreicht. Das bedeutet, dass die Schätzgewichte der teilnehmenden Schüler*innen weitgehend übereinstimmen, da sie sich durch Multiplikation der inversen Ziehungswahrscheinlichkeiten ergeben. Homogene Gewichte erhöhen in der Regel die Präzision der Schätzungen. Für Sekundärpopulationen wie die Schulen selbst sind die Schätzgewichte aufgrund der variierenden Ziehungswahrscheinlichkeiten entsprechend variabler (Meinck, 2020).

Im Rahmen einer Small-School-Analyse wurden kleine Schulen im Sampling Frame gesondert berücksichtigt, da eine hohe Anzahl kleiner Schulen die Repräsentativität der Schüler*innenstichprobe sowie die praktische Durchführung der Studie beeinflussen kann. Die

Analyse diente der Prüfung des Umgangs mit sehr kleinen Schulen im Rahmen der Stichprobenziehung.

Für den Fall, dass eine gezogene Schule nicht an der Studie teilnehmen konnte, wurden bereits bei der Ziehung Ersatzschulen festgelegt. Die direkt vor und nach einer gezogenen Schule im sortierten Sampling Frame liegenden Schulen wurden als erste und zweite Ersatzschule bestimmt. Durch Sortierung nach den impliziten Stratifizierungsvariablen und der MOS wurde sichergestellt, dass die Ersatzschulen der Originalschule in ihren Merkmalen möglichst ähnlich waren, sodass sie diese im Bedarfsfall hinreichend zuverlässig ersetzen konnten.

Die Kombination aus expliziter und impliziter Stratifizierung, der systematischen Sortierung des Sampling Frames und der Anwendung des PPS-Verfahrens im Ziehungsverfahren gewährleistet, dass die Stichprobe ein möglichst genaues und unverzerrtes Abbild der Schulen liefert, die 15-jährige Schüler*innen unterrichten.

Für die deutsche PISA-2025-Schulstichprobe wurden 262 Schulen gezogen. Für jede gezogene Schule wurden zwei potenzielle Ersatzschulen identifiziert.

Realisierte Schulstichprobe

In einem ersten Schritt wurden die original gezogenen Schulen kontaktiert und ihr Teilnahmestatus geklärt. Dabei galt, dass die Teilnahme aller öffentlichen Schulen verpflichtend war und für Schulen in freier Trägerschaft je nach Bundesland entweder verpflichtend oder freiwillig (s. Tabelle 14.2). Als Ergebnis der Kontaktaufnahme konnten verschiedene Fälle eintreten: 1. die Schule bestätigte ihre Teilnahme, 2. die Schule sagte die Teilnahme ab oder konnte aus bestimmten Gründen nicht teilnehmen (z. B. aufgrund außergewöhnlicher Ereignisse wie einem Brand) oder 3. die Schule wurde als nicht teilnahmeberechtigt eingestuft. Ausschlussgründe für eine Teilnahmeberechtigung waren das Fehlen PISA-teilnahmeberechtigter Schüler*innen, eine Schulschließung oder der Umstand, dass an der Schule ausschließlich Schüler*innen der nationalen Option der gezogenen 9. Klasse vertreten waren, jedoch keine Schüler*innen der PISA-Kernpopulation. Diese Ausschlussgründe waren Teil der Stichprobenlogik und spiegelten die Verteilung solcher Schulen in der Grundgesamtheit wider, sodass in diesen Fällen auch keine Ersatzschule kontaktiert wurde. Sagte jedoch eine Originalschule die Teilnahme ab oder konnte aus anderen Gründen nicht teilnehmen, obwohl grundsätzlich teilnahmeberechtigte Schüler*innen vorhanden waren, wurde dies als Nichtteilnahme gewertet und die erste Ersatzschule kontaktiert. Bei deren Nichtteilnahme wurde die zweite Ersatzschule kontaktiert. Wenn weder die Originalschule noch die zugeordneten Ersatzschulen an der Studie teilnahmen, führte dies zum Ausfall der entsprechenden Stichprobeneinheit.

Unabhängig vom im Zuge der Schulrekrutierung festgestellten Teilnahmestatus wurden Schulen auch nach Durchführung der Testung als nichtteilnehmend eingestuft, wenn die Teilnahmequote der ausgewählten, teilnahmeberechtigten und nicht ausgeschlossenen Schüler*innen bei 33 Prozent oder weniger lag. In diesem Fall wurden Daten dieser Schulen in den Analysen nicht berücksichtigt. Die Teilnahmequote bezog sich ausschließlich auf die ausgewählten Schüler*innen der PISA-Kernpopulation. In Schulen mit nationaler Option wurde sie getrennt für die PISA-Kernpopulation und die zusätzliche Stichprobe bestimmt, sodass eine Schule für die nationale Option der Neuntklässler*innen als teilnehmend gelten konnte, während sie für PISA als nichtteilnehmend eingestuft wurde.

Die ursprünglich gezogene Brutto-Schulstichprobe umfasste 262 Schulen. Diese verminderte sich um neun Schulen, von denen acht keine PISA-teilnahmeberechtigten Schüler*innen hatten (davon eine Schule mit mehreren Bildungsgängen, eine Schule nicht-deutscher Herkunftssprache sowie sechs Berufsschulen) und eine geschlossen worden war (eine Berufsschule). Diese Schulen wurden gemäß dem PISA-Protokoll nicht ersetzt. Die korrigierte

Bruttostichprobe umfasste daher 253 Schulen. Sechs Schulen verweigerten die Teilnahme und konnten durch keine ihrer beiden Ersatzschulen ersetzt werden (davon eine integrierte Gesamtschule, eine Hauptschule sowie vier Berufsschulen). Eine Schule (Gymnasium) erreichte nicht die erforderliche Teilnahmequote von mehr als 33 Prozent der Schüler*innen und wurde daher nachträglich als nicht teilnehmend eingestuft. Insgesamt konnte somit eine Stichprobe von 246 Schulen realisiert werden, darunter elf erste Ersatzschulen und sechs zweite Ersatzschulen.

Die ungewichtete Teilnahmequote auf Schulebene lag bezogen auf die original gezogenen Schulen bei 90,5 Prozent und erhöhte sich nach Einbezug der ersten Ersatzschulen auf 94,9 Prozent sowie nach Einbezug der zweiten Ersatzschulen auf 97,2 Prozent. Die gewichtete Teilnahmequote auf Schulebene betrug 93,7 Prozent ohne Berücksichtigung von Ersatzschulen, 97 Prozent nach Einbezug der ersten Ersatzschulen und 99,1 Prozent nach Einbezug der zweiten Ersatzschulen. Damit wurde die international geforderte Mindestteilnahmequote der original gezogenen Schulen von 85 Prozent deutlich überschritten, sodass die Stichprobe den internationalen Qualitätsanforderungen entspricht.

Ziehung der Schüler*innenstichproben

Für die Listungen und den Informationsaustausch mit den Schulen kam die Webanwendung IEA OnlineSurveyExpert (IEA OSE) zum Einsatz, über die die Schulen alle ziehungsrelevanten Personen samt den relevanten Daten erfassten. Die IEA Hamburg erhielt diese Listen ausschließlich in pseudonymisierter Form. Den erfassten Personen wurden Ordnungsnummern zugewiesen, die auch in den von den Schulen erstellten Personenlisten enthalten waren und die Grundlage für die Kommunikation zwischen Schulen und IEA Hamburg bildeten.[4]

Um sicherzustellen, dass die Schüler*innenstichproben den Anforderungen des Studiendesigns entsprachen, wurden die teilnehmenden Schulen zunächst gebeten, eine Liste aller potenziell teilnahmeberechtigten Schüler*innen zu erstellen. Über IEA OSE wurden hierfür alle 15-Jährigen und Neuntklässler*innen erfasst, einschließlich Angaben zu Geschlecht, Klassenstufe, Klassenbezeichnung, Geburtsmonat und -jahr, Bildungsgang sowie einem möglichen sonderpädagogischen Förderbedarf (SPF).

Für die Ziehung der PISA-Kern- und PISA-FLA-Stichproben sowie der Stichprobe der Neuntklässler*innen wurde die vom internationalen Konsortium bereitgestellte Maple-Software eingesetzt. Auf Grundlage der von den Schulen erstellten Schüler*innenlisten wurde zunächst eine Liste aller 9. Klassen generiert, einschließlich der Anzahl der jeweiligen Schüler*innen. Diese Klassenlisten wurden in einem ersten Schritt in Maple eingelesen und eine 9. Klasse pro Schule zufällig gezogen (s. Abschnitt *Stichprobendesign*). In einem zweiten Schritt wurden die vollständigen Schüler*innenlisten um die Kennzeichnung der Zugehörigkeit zur gezogenen 9. Klasse ergänzt und anschließend in Maple importiert. Abschließend erfolgte die Ziehung der PISA-Kern- und PISA-FLA-Stichproben sowie der Stichprobe der Neuntklässler*innen. Detaillierte Beschreibungen der jeweiligen Stichproben finden sich in den Abschnitten *Stichprobe der 15-Jährigen* und *Stichprobe der Neuntklässler*innen*. Die in Tabelle 14.4 dargestellten Zahlen zeigen die Anzahl der gezogenen Schüler*innen (Bruttostichproben) in den jeweiligen Stichproben sowie deren Verteilung auf die verschiedenen Schularten.

4 In IEA OSE werden personenbezogene Angaben passwortgeschützt erfasst und verschlüsselt übertragen, sodass insbesondere Namensangaben nur durch die jeweilige Schule eingesehen werden können. Die weitere Verarbeitung erfolgte auf Basis automatisch vergebener Ordnungsnummern, die eine Verknüpfung der Daten ohne Zugriff auf Klarnamen ermöglichen.

Tabelle 14.4: Anzahl der gezogenen Schüler*innen in den Untersuchungsstichproben und deren Verteilung auf die Schularten

Stichproben	HS	RS	MB	GY	IG	ND	∑allg	FS	BS	∑
PISA-Kernstichprobe	564	1.400	943	2.803	1659	13	**7.382**	159	156	7.697
PISA-FLA-Stichprobe	168	425	284	836	491	4	**2.208**		46	2.254
Stichprobe der Neuntklässler*innen	247	690	413	1.408	763	1	**3.522**	98		3.620
∑	**979**	**2.515**	**1640**	**5.047**	**2.913**	**18**	**13.112**	**257**	**202**	**13.571**

Anmerkung: ∑allg: allgemeine Schulen; ∑: alle Schulen. HS = Hauptschule, RS = Realschule, MB = Schule mit mehreren Bildungsgängen, GY = Gymnasium, IG = integrierte Gesamtschule, ND = Schule nicht-deutscher Herkunftssprache, FS = Förderschule, BS = Berufsschule.

Realisierte Schüler*innenstichproben

Im Folgenden wird die Entwicklung von den im Abschnitt *Ziehung der Schüler*innenstichproben* beschriebenen Bruttostichproben zu den bereinigten und realisierten Stichproben dargestellt. Im Rahmen der Bereinigung wurden zunächst Schüler*innen aus der Stichprobe entfernt, deren Schulen aufgrund einer Beteiligungsquote von 33 Prozent oder weniger nachträglich als nicht teilnehmend klassifiziert wurden (s. Abschnitt *Realisierte Schulstichprobe*). Dies betraf lediglich eine Schule. Darüber hinaus wurden Schulabgänge und Ausschlüsse aus der Stichprobe entfernt. Als Schulabgänge galten Schüler*innen, die nach der Ziehung, aber vor der Testdurchführung die Schule verlassen hatten. Zu den Ausschlüssen zählten Schüler*innen, die aufgrund eines sonderpädagogischen Förderbedarfs oder wegen einer Beschulungsdauer von unter einem Jahr im nationalen Bildungssystem nicht teilnahmen (OECD, 2026a). Die verbleibende bereinigte Stichprobe umfasste ausschließlich teilnahmeberechtigte Schüler*innen. Nach internationalen Vorgaben galten Schüler*innen als teilnehmend, wenn sie mindestens fünf Aufgaben des kognitiven Tests bearbeitet hatten oder mindestens eine Aufgabe des kognitiven Tests sowie alle für die Berechnung des sozioökonomischen Index erforderlichen Fragen des Schüler*innenfragebogens beantwortet hatten (OECD, 2026a). Nichtteilnahmen konnten demnach sowohl auf die Nichterfüllung dieser Anforderungen als auch auf andere Gründe, beispielsweise krankheitsbedingte Abwesenheiten, zurückzuführen sein.

Die bereinigte Stichprobe umfasste insgesamt 13.117 Schüler*innen und setzte sich aus der PISA-Kern- und der PISA-FLA-Stichprobe sowie der Stichprobe der Neuntklässler*innen zusammen, die jeweils eigenständige Stichproben bildeten. Tabelle 14.5 gibt einen Überblick über die Verteilung der Schüler*innen nach Teilnahmestatus bei den kognitiven Testinstrumenten und dem Schüler*innenfragebogen in den einzelnen Untersuchungsstichproben sowie über die jeweilige Gesamtzahl der Schüler*innen der bereinigten Stichprobe. Darüber hinaus ist die Anzahl der Schulabgänge, die Anzahl der Ausschlüsse sowie die Anzahl der Schüler*innen aus der Schule mit einer Beteiligung von höchstens 33 Prozent separat ausgewiesen.

Tabelle 14.5: Anzahl der Schüler*innen nach Teilnahmestatus am kognitiven Test (Teilnahmen und Nichtteilnahmen) in den Untersuchungspopulationen sowie die Gesamtzahl der in die Auswertung einbezogenen Schüler*innen. Zusätzlich ausgewiesen sind die Anzahl der Schulabgänge, der Ausschlüsse sowie die Anzahl der Schüler*innen in Schulen mit einer Beteiligung von ≤ 33 Prozent, die in den Auswertungen unberücksichtigt bleiben.

Stichproben	Teilnahmen	Nichtteilnahmen	∑	Schulabgänge	Ausschlüsse	Beteiligung ≤ 33 %
PISA-Kernstichpropbe	**6.659**	766	**7.425**	145	113	14
PISA-FLA-Stichprobe	**1.865**	310	**2.175**	42	32	5
Stichprobe der Neuntklässler*innen	**3.184**	333	**3.517**	49	42	12
∑	**11.407**	1.710	**13.117**	236	187	31

Die Teilnahmequote auf Ebene der Schüler*innen für die kognitiven Testinstrumente der PISA-Kernstichprobe lag ungewichtet bei 89,7 Prozent bezüglich der ursprünglich gezogenen Schulen und bei 89,7 Prozent nach Einbezug der Ersatzschulen. Die gewichtete Teilnahmequote betrug 89,3 Prozent in den original gezogenen Schulen und 89,3 Prozent nach Einbezug der Ersatzschulen. Für PISA gilt eine Mindestanforderung von 80 Prozent für die gewichtete Teilnahmequote der aus der PISA-Kernpopulation gezogenen Schüler*innen in den teilnehmenden Schulen (OECD, 2026a). Da diese Anforderung erfüllt wurde, kann die realisierte Stichprobe als hinreichend repräsentativ und methodisch belastbar angesehen werden.

Für die PISA-FLA-Stichprobe betrug die ungewichtete Teilnahmequote für die kognitiven Testinstrumente unter Einbezug der Ersatzschulen 85,7 Prozent. Die entsprechende gewichtete Teilnahmequote lag bei 85,9 Prozent. Auch für die PISA-FLA-Stichprobe wurde damit die international festgelegte Mindestanforderung einer gewichteten Teilnahmequote von mindestens 80 Prozent für die aus der FLA-Zielpopulation gezogenen Schüler*innen in den teilnehmenden Schulen mit FLA-teilnahmeberechtigten Schüler*innen erfüllt (OECD, 2026a).

In der Stichprobe der Neuntklässler*innen lag die ungewichtete Teilnahmequote für die kognitiven Testinstrumente bei 90,5 Prozent nach Einbezug der Ersatzschulen. Da diese nationale Option kein Bestandteil der internationalen PISA-Erhebung war, liegt für diese Stichprobe keine international berichtete Teilnahmequote vor. Der Datensatz der Neunklässler*innen umfasst insgesamt 4.142 Schüler*innen, wovon 958 für die PISA-Kernstichprobe gezogen wurden (s. Abschnitt Stichprobe der Neuntklässler*innen).

Für alle gezogenen Schüler*innen wurde jeweils ein Elternfragebogen ausgegeben. Berücksichtigt wurden jedoch ausschließlich Elternfragebögen der Eltern von Schüler*innen, die an der Studie teilgenommen hatten. Bezogen auf die 11.407 teilnehmenden Schüler*innen entspricht dies einer ungewichteten Teilnahmequote von 42,6 Prozent. Eine getrennte Berichterstattung nach den einzelnen Schüler*innenstichproben ist nicht möglich, da hierfür keine stichprobenspezifischen Teilnahmezahlen vorliegen.

Festlegung der Lehrkräftestichproben

Für PISA 2025 wurden an den teilnehmenden Schulen, die im Abschnitt *Stichprobe der Lehrkräfte* definierten Lehrkräfte befragt. Die Schulen wurden gebeten, alle infrage kommenden Lehrkräfte zu listen und deren Unterrichtsfächer entsprechend zu kennzeichnen. Lehrkräfte, die sowohl Englisch als auch ein naturwissenschaftliches Fach unterrichte-

ten, wurden in beiden Kategorien erfasst und erhielten einen Fragebogen, der sowohl Naturwissenschaften als auch Englisch umfasste. Die Lehrkräftelisten wurden anschließend mittels IEA OSE an die IEA Hamburg in pseudonymisierter Form übermittelt und enthielten lediglich Angaben zu den Unterrichtsfächern. Die Listen wurden in Maple importiert und die Lehrkräftestichproben wurden aus allen gelisteten Lehrkräften gezogen. Insgesamt wurden 4.887 Lehrkräfte erfasst. Davon unterrichteten 1.984 mindestens ein naturwissenschaftliches Fach, 1.236 Englisch und 155 sowohl Englisch als auch mindestens ein naturwissenschaftliches Fach. Letztere waren Bestandteil sowohl der Stichprobe der naturwissenschaftlich unterrichtenden Lehrkräfte als auch der Stichprobe der Englischlehrkräfte. Weitere 1.512 Englischlehrkräfte wurden im Rahmen der Begleitstudie PISA-Schreiben befragt. Die Bruttostichproben umfassten damit 2.139 Fälle der Stichprobe der naturwissenschaftlich unterrichtenden Lehrkräfte, 1.391 der Stichprobe der Englischlehrkräfte und 1.512 der Begleitstudie PISA-Schreiben. Insgesamt wurden 5.042 Stichprobenfälle gebildet.

Realisierte Lehrkräftestichproben

Ausgehend von den zuvor gebildeten Bruttostichproben der Lehrkräfte wurden diese zunächst um Lehrkräfte bereinigt, die die Schule nach der Listung und vor der Befragung verlassen hatten, bei denen ein Listungsfehler vorlag oder die an einer Schule mit zu niedriger Schüler*innenbeteiligung unterrichteten. Nach der Bereinigung verblieben insgesamt 4.809 Lehrkräfte in den Stichproben. Lehrkräfte, die sowohl Englisch als auch mindestens ein naturwissenschaftliches Fach unterrichteten, beantworteten einen gemeinsamen Fragebogen mit einem naturwissenschaftlichen und einem englischbezogenen Teil. Für die Auswertungen wurde jeweils ausschließlich der entsprechende fachbezogene Teil des Fragebogens der jeweiligen Lehrkräftestichprobe zugeordnet. Als teilnehmend galten Lehrkräfte, die mindestens eine gültige Antwort im Fragebogen gegeben hatten (OECD, 2026a). Für Lehrkräfte wurden keine Gewichtungsfaktoren berechnet, daher werden ausschließlich ungewichtete Teilnahmequoten berichtet.

Die bereinigte Stichprobe der naturwissenschaftlichen Lehrkräfte umfasste 2.111 Lehrkräfte. Davon unterrichteten 1.958 Lehrkräfte ausschließlich naturwissenschaftliche Fächer und 153 mindestens ein naturwissenschaftliches Fach sowie Englisch. Insgesamt nahmen 1.691 Lehrkräfte teil, darunter 1.568 Lehrkräfte, die ausschließlich naturwissenschaftliche Fächer unterrichteten, sowie 123 Lehrkräfte, die zusätzlich Englisch unterrichteten. Dies entspricht einer ungewichteten Teilnahmequote von 80,1 Prozent.

Die Stichproben der Englischlehrkräfte umfassten nach der Bereinigung insgesamt 2.851 Lehrkräfte. Davon unterrichteten 1.227 Lehrkräfte ausschließlich Englisch und 153 Englisch sowie mindestens ein naturwissenschaftliches Fach. Die Begleitstudie PISA-Schreiben umfasste weitere 1.471 Englischlehrkräfte. Für die Befragung der Englischlehrkräfte und die Begleitstudie PISA-Schreiben liegen keine getrennten Teilnahmezahlen vor, eine getrennte Berichterstattung ist daher nicht möglich. Insgesamt nahmen 2.190 Englischlehrkräfte teil. Bezogen auf die gemeinsame bereinigte Stichprobe entspricht dies einer ungewichteten Teilnahmequote von 76,8 Prozent.

14.3.3 Gewichtung

Zur Analyse der PISA-Daten werden aus Gewichtungsfaktoren abgeleitete Schätzgewichte eingesetzt. Dies ermöglicht trotz des komplexen Stichprobendesigns unverzerrte und für die Zielpopulation repräsentative Schätzungen von Populationsparametern (z. B. Schätzungen von mittleren Kompetenzwerten auf Ebene der teilnehmenden Staaten). Da die Auswahlwahrscheinlichkeiten der Schüler*innen variierten und nicht alle gezogenen Einheiten

an der Erhebung teilnahmen, stellen die Gewichte sicher, dass jede teilnehmende Person die korrekte Anzahl von Schüler*innen in der Gesamtpopulation repräsentiert. So werden Unterschiede in den Auswahlwahrscheinlichkeiten ausgeglichen, indem Schüler*innen mit einer geringeren Auswahlwahrscheinlichkeit ein höheres Gewicht erhalten als Schüler*innen mit einer höheren Auswahlwahrscheinlichkeit. Die Berücksichtigung der Schätzgewichte wirkt sich auch auf die Berechnung von Teilnahmequoten aus. Ungewichtete Teilnahmequoten berücksichtigen alle teilnehmenden und nicht teilnehmenden Schulen beziehungsweise Schüler*innen unabhängig von ihrer Auswahlwahrscheinlichkeit in gleichem Maße. Gewichtete Teilnahmequoten berücksichtigen zusätzlich die unterschiedlichen Auswahlwahrscheinlichkeiten der Einheiten. Unterschiede zwischen gewichteten und ungewichteten Teilnahmequoten entstehen insbesondere dann, wenn sich die Teilnahme auf Schul- oder Schüler*innenebene mit hohen und niedrigen Gewichten unterscheidet. Insgesamt folgten die angewendeten Verfahren den etablierten Standards zur Analyse komplexer Stichprobendaten und gewährleisteten eine angemessene Berücksichtigung von Ziehungswahrscheinlichkeiten und Teilnahmemustern (OECD, 2024). Die entsprechende Dokumentation für PISA 2025 erfolgt im zugehörigen Technical Report (OECD, 2026b).

14.3.4 Präzision der PISA-Daten

Alle berichteten Ergebnisse der PISA-Studie enthalten jeweils eine Angabe zur Präzision der geschätzten Populationsparameter in Form des ▶ Standardfehlers (SE). Die Präzision der PISA-Ergebnisse ergibt sich aus dem Zusammenspiel zweier methodischer Komponenten, dem Stichprobendesign und dem Assessmentdesign.

Zur Berücksichtigung unterschiedlicher Auswahlwahrscheinlichkeiten und der Nichtteilnahme werden die soeben im Abschnitt *Gewichtung* beschriebenen Stichprobengewichte verwendet. Diese ermöglichen die Schätzung unverzerrter Populationsparameter (z. B. ▶ Mittelwerte, Anteile von Schüler*innen in bestimmten ▶ Kompetenzstufen sowie Gruppenunterschiede).

Das komplexe Stichprobendesign hat jedoch ebenfalls einen Einfluss auf den Stichprobenfehler: Durch die zweistufige Ziehung entsteht eine hierarchische (das heißt geschachtelte) Datenstruktur, in der Schüler*innen jeweils einer Schule zugeordnet sind. Schüler*innen innerhalb einer Schule sind sich womöglich ähnlicher als solche aus unterschiedlichen Schulen. Mit anderen Worten sind Beobachtungen innerhalb derselben Schule nicht unabhängig voneinander – dieser Effekt wird als Designeffekt bezeichnet. Im Vergleich zu einer einfachen Zufallsstichprobe ist eine solche Stichprobe deutlich weniger präzise und weist daher einen signifikant größeren Stichprobenfehler auf. Zur korrekten Berücksichtigung dieses komplexen Designs werden bei PISA Replikatgewichte zur Schätzung der Standardfehler eingesetzt. Diese Replikatgewichte werden mittels Balanced Repeated Replication (BRR) mit Fay-Anpassung ermittelt (OECD, 2024). Dieses Verfahren gewährleistet die korrekte Schätzung der Standardfehler und verhindert eine Unterschätzung der Unsicherheit. Eine zusätzliche Quelle statistischer Unsicherheit, der Messfehler aus den Kompetenzschätzungen, wird ebenfalls bei der Schätzung der Standardfehler berücksichtigt und ist in Abschnitt 14.5 zur Messung und Skalierung genauer beschrieben.

14.4 Testorganisation und Durchführung

Carola Bretsch, Jennifer Diedrich & Tamara Kastorff

In diesem Abschnitt werden die organisatorischen und praktischen Aspekte der Umsetzung der PISA-Studie 2025 in Deutschland beschrieben. Mit der Feldarbeit wird seit PISA 2000 die IEA Hamburg von der nationalen Studienleitung (ZIB an der TUM) beauftragt. Hier werden die Tätigkeiten der IEA Hamburg in der Vorbereitung und Durchführung an den teilnehmenden Schulen in Deutschland dargestellt. Darüber hinaus wird anhand eines exemplarischen Ablaufs ein typischer PISA-Testtag an den Schulen veranschaulicht.

14.4.1 Beteiligte Akteur*innen

An der Vorbereitung und Durchführung der Testsitzungen waren unterschiedliche Akteur*innen beteiligt, deren Aufgaben klar voneinander abgegrenzt waren. Während die nationale Studienleitung die wissenschaftliche Organisation, Konzeption und Begleitung der Studie verantwortete, übernahm die IEA Hamburg als nationaler Auftragnehmer die Feldarbeit sowie zentrale operative Aufgaben im Datenmanagement. Die Durchführung erfolgte in enger Zusammenarbeit mit den Schulkoordinator*innen, die die organisatorische Vorbereitung an den Schulen verantworteten, sowie den von der IEA Hamburg eingesetzten Testleiter*innen, die die Testsitzungen vor Ort durchführten. Die einzelnen Aufgaben der Akteur*innen werden nachfolgend kurz beschrieben.

Die IEA Hamburg war mit der zentralen Organisation und Durchführung der Studie betraut. Die Abteilung Feldarbeit fungierte dabei als Schnittstelle zwischen der nationalen Studienleitung, den Bildungs- beziehungsweise Kultusministerien und den an der PISA-Studie teilnehmenden Schulen. Zu ihren Aufgaben zählten die Kommunikation und kontinuierliche Betreuung der teilnehmenden Schulen über den gesamten Projektverlauf hinweg. Dies umfasste die Erstellung und Übersetzung von Manualen, Skripten, Anleitungen und Anschreiben sowie die Rekrutierung, Schulung und Einsatzkoordination der Testleiter*innen sowie die Organisation und den Versand der Testmaterialien.

Zur Unterstützung der Schulen führte die IEA Hamburg, unter Beteiligung des ZIB, mehrere bundeslandübergreifende Online-Informationsveranstaltungen durch. Ziel dieses Angebots war es, Schulleitungen und Schulkoordinator*innen umfassend über Ziele, Ablauf und Anforderungen der PISA-Studie zu informieren sowie offene Fragen zu klären. Schätzungsweise 90 Prozent der teilnehmenden Schulen nahmen an einer der angebotenen Veranstaltungen teil.

Jede an der PISA-Studie teilnehmende Schule benannte eine*n Schulkoordinator*in, der beziehungsweise die im Vorfeld der Testsitzungen alle organisatorischen und administrativen Aufgaben an der Schule verantwortete. Die Schulkoordinator*innen fungierten als zentrale Ansprechpersonen für die IEA Hamburg und deren Testleiter*innen. Zu ihren wesentlichen Aufgaben gehörten vorbereitende Dateneingaben in der Webanwendung IEA OSE, insbesondere die vollständige Erfassung aller zur Zielstichprobe gehörenden Schüler*innen und Lehrkräfte, sowie sämtliche Abstimmungsprozesse mit allen Beteiligten innerhalb der Schule.

Die Testsitzungen wurden von sorgfältig ausgewählten und qualifizierten Testleiter*innen durchgeführt. Diese wurden in einem mehrstufigen Schulungsprozess, der unter anderem mehrstündige Präsenzschulungen umfasste, von der IEA Hamburg intensiv auf eine standardisierte Durchführung der PISA-Erhebung vorbereitet. Ergänzend erhielten sie schriftliche Handreichungen, darunter ein detailliertes Testleitungsmanual sowie ein

standardisiertes Testleitungsskript mit verbindlichen Durchführungsvorgaben. Zudem dokumentierten sie den Verlauf jeder Testsitzung in standardisierten Protokollen. So wurde gewährleistet, dass alle Testleiter*innen gleichermaßen gut vorbereitet waren und die Erhebungen unter vergleichbaren Bedingungen durchgeführt werden konnten. Während des gesamten Testfensters wurden die Testleiter*innen durch die Abteilung Feldarbeit der IEA Hamburg betreut und standen in engem Austausch mit den Schulkoordinator*innen. Die Kombination aus qualifizierter Auswahl, intensiver Schulung, verbindlichen Durchführungsvorgaben, fortlaufender Betreuung und systematischer Dokumentation trug wesentlich zur Qualitätssicherung und Vergleichbarkeit der erhobenen Daten bei.

Zur Überprüfung der Einhaltung der internationalen Durchführungsstandards wurden die Testsitzungen gemäß den Technical Standards der OECD (2026a) im Rahmen des sogenannten PISA Quality Monitoring in allen Teilnahmestaaten stichprobenartig kontrolliert. Hierfür wurden unabhängige, speziell geschulte und von der internationalen Studienleitung vergütete pädagogische Fachkräfte eingesetzt. Diese nahmen an den Testleitungsschulungen teil, beobachteten zufällig ausgewählte Testsitzungen hinsichtlich der Einhaltung der international abgestimmten Manuale und Skripte und berichteten ihre Beobachtungen an die internationale Studienleitung. Für PISA 2025 wurden in Deutschland keine Beanstandungen hinsichtlich der Durchführung der Studie festgestellt.

14.4.2 Technische Vorbereitung und Systemcheck

Die Testaufgaben und Fragebögen wurden im Rahmen von PISA 2025 online erhoben. Die Durchführung erfolgte entweder über die schulischen Geräte und die Internetverbindung der Schule oder, sofern die technischen Voraussetzungen nicht vollständig gegeben waren, über die von der IEA Hamburg bereitgestellten Laptops und mobilen Hotspots, die von den Testleiter*innen zum Testtermin mitgebracht wurden. Die Schüler*innen für die Englischtestung (s. Abschnitt 14.3.1 *Stichprobe der 15-Jährigen*) wurden zusätzlich von den Testleiter*innen mit den erforderlichen Headsets ausgestattet.

Zur Sicherstellung der technischen Voraussetzungen wurde im Vorfeld der Testsitzungen an den Schulen ein Systemcheck durchgeführt. Schulen mit grundsätzlich geeigneter technischer Ausstattung prüften dabei mithilfe einer von der internationalen Studienleitung bereitgestellten Prüfroutine die Kompatibilität ihrer Systeme mit den Anforderungen der onlinebasierten Erhebung. Die Durchführung erfolgte durch den/die Schulkoordinator*in oder eine in der Schule für die IT zuständige Person.

Im Rahmen des Systemchecks wurden technische Parameter wie Betriebssystem, Bildschirmgröße, Browser und Mikrofon überprüft. Ergänzend wurde die Internetgeschwindigkeit (Upload und Download) gemessen sowie ein Soundcheck der in der FLA-Testgruppe eingesetzten Kopfhörer durchgeführt. Insgesamt konnte in PISA 2025 an ca. 60 Prozent aller Schulen die schuleigene Infrastruktur für die Testdurchführung genutzt werden.

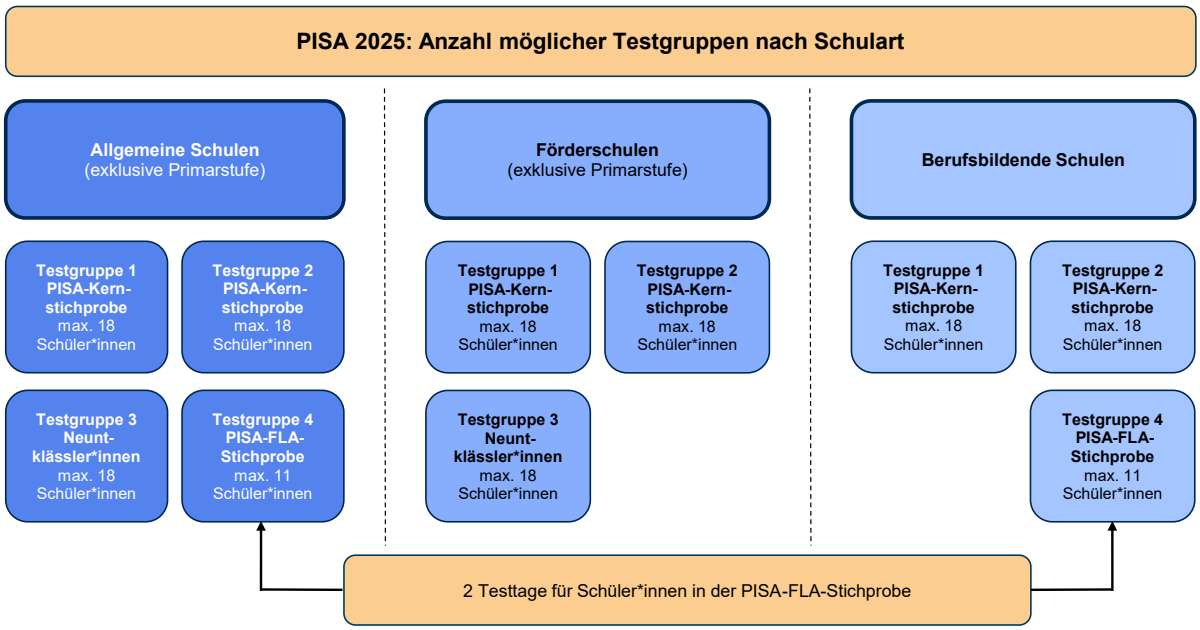
14.4.3 Zusammensetzung der Testgruppen und Durchführung der Testsitzungen

An allgemeinbildenden Schulen (mit Ausnahme von Förderschulen) war vorgesehen, dass klassenübergreifend bis zu 63 Schüler*innen an PISA 2025 teilnehmen. Aus der Gesamtheit der 15-Jährigen einer Schule wurden durch die IEA Hamburg gemäß den internationalen Regeln per Zufallsziehung bis zu 45 Schüler*innen zur Teilnahme ausgewählt und auf drei Testgruppen verteilt (zwei Testgruppen der PISA-Kernstichprobe sowie eine Testgruppe der

PISA-FLA-Stichprobe). Zusätzlich wurde aus allen 9. Klassen eine Klasse gezogen, aus der altersunabhängig bis zu 18 Schüler*innen für eine weitere Testgruppe (nationale Option der Modalklasse) ausgewählt wurden (s. Abschnitt 14.3.1 *Stichprobe der Neuntklässler*innen*). Daraus ergaben sich an allgemeinen Schulen in der Regel vier Testgruppen mit folgender typischer Zusammensetzung (siehe Abbildung 14.5): zwei Testgruppen mit jeweils bis zu 18 Schüler*innen der PISA-Kernstichprobe der 15-Jährigen (Testgruppen 1 und 2), eine Testgruppe mit bis zu 18 Schüler*innen aus der gezogenen 9. Klasse (Testgruppe 3) sowie eine Testgruppe der PISA-FLA-Stichprobe mit maximal 11 Schüler*innen (Testgruppe 4). Im Unterschied zu den ersten drei Testgruppen hatte Testgruppe 4 (PISA-FLA-Stichprobe) zwei Testsitzungen an zwei aufeinanderfolgenden Tagen (s. Abschnitt 14.2.3 *Nationale Ergänzungen*). An Förderschulen und berufsbildenden Schulen reduzierte sich die maximale Anzahl auf drei Testgruppen, da entweder die PISA-FLA-Stichprobe (Förderschulen) oder die Stichprobe der Neuntklässler*innen (berufsbildende Schulen) entfiel.

Die Testdurchführung erfolgte in separaten Räumen je Testgruppe. Abhängig von der technischen und räumlichen Ausstattung der Schulen erstreckte sich die Durchführung aller Testgruppen über zwei bis fünf Tage. Die Möglichkeit zur parallelen Durchführung mehrerer Testsitzungen unter Nutzung der Schulhardware hing maßgeblich von der Anzahl verfügbarer Geräte und Räume ab, war aber grundsätzlich möglich. Standen nicht ausreichend viele Geräte oder Räume zur Verfügung, wurden die Testsitzungen an verschiedenen Tagen durchgeführt. Wurden hingegen Laptops der IEA Hamburg bereitgestellt, fanden die Testsitzungen in der Regel parallel für alle Testgruppen statt. Unabhängig davon waren an allen Schulen (mit Ausnahme der Förderschulen) mindestens zwei Termine erforderlich, da die PISA-FLA-Testgruppe an zwei, in der Regel aufeinanderfolgenden Tagen getestet wurde.

Abbildung 14.5: Schematische Darstellung der Anzahl möglicher Testgruppen an den Schulen nach Schularten mit Schüler*innenstichproben (PISA-Kernstichprobe, Stichprobe der Zusatzdomäne FLA und nationale Option der Neuntklässler*innen) sowie Anzahl Schüler*innen pro Testgruppe



14.4.4 Exemplarischer Ablauf eines PISA-Testtages Testgruppe 1–3 (Kernstichprobe und nationale Option)

Alle PISA-Testsitzungen fanden vormittags in den Schulen statt. Am Testtag trafen sich die Testleiter*innen rund eine Stunde vor Testbeginn mit dem/der Schulkoordinator*in, um letzte Absprachen zu treffen und die Testräume vorzubereiten. Jede Testsitzung wurde von den Testleitungen in einem separaten Raum durchgeführt. Zur Einhaltung der schulischen Aufsichtspflicht stellte die Schule für jede Testsitzung eine Aufsichtsperson aus dem Kollegium, die während des gesamten Verlaufs der Testsitzung im Raum anwesend war. Eine schematische Darstellung des nachfolgend beschriebenen Ablaufs findet sich in Abbildung 14.6.

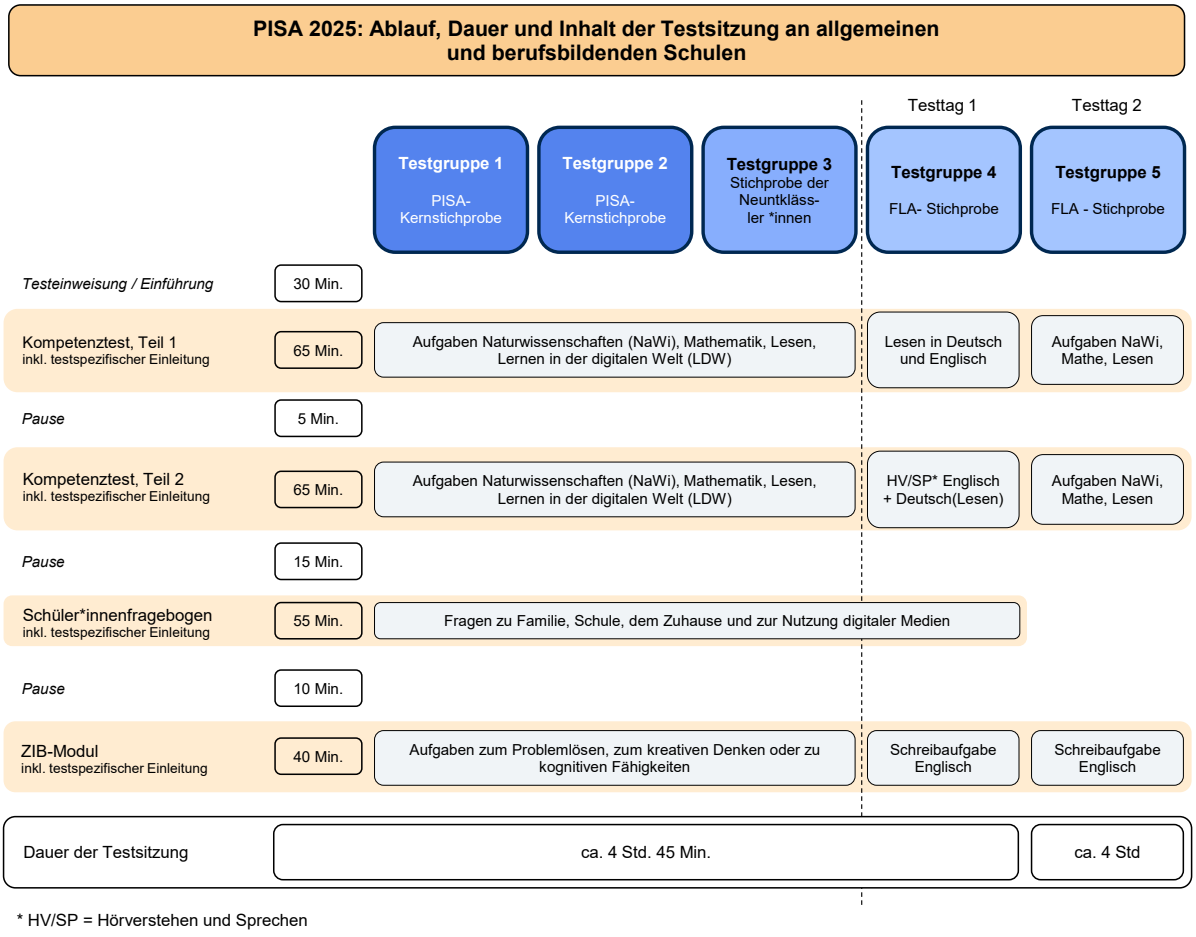
Zur Sicherstellung der Pseudonymität der Schüler*innen erhielten Schulkoordination und Testleitung unterschiedliche Auszüge aus der Webanwendung IEA OSE, welche ausschließlich von der Schulkoordination verknüpft werden konnten. Am Ende des Testfensters und vor der Übergabe der Daten an die internationale Studienleitung wurde diese Verknüpfung gelöscht, sodass die PISA-Daten ab diesem Zeitpunkt anonym sind.

In Bundesländern, in denen der Schüler*innenfragebogen freiwillig ist (s. Tabelle 14.2), überprüften die Testleiter*innen am Testmorgen zusätzlich die Einwilligungsblätter der Eltern. Vor Beginn der Testsitzung richteten die Testleiter*innen die Arbeitsplätze ein und wiesen die Testinstrumente schülerindividuell zu. Zudem wurden die Elternfragebögen inkl. Rücksendeumschlag und weitere Informationen verteilt. Die ausgefüllten Fragebögen wurden in der Schule gesammelt und anschließend an die IEA Hamburg zurückgesendet.

Nach Abschluss der Vorbereitungen nahmen die Schüler*innen die ihnen zugewiesenen Arbeitsplätze ein. Die Bearbeitung der Testinstrumente begann für alle Schüler*innen gleichzeitig unter Anleitung der Testleiter*innen. An allen Schulen mit Ausnahme der Förderschulen umfasste die Testsitzung zwei Kompetenztests mit einer Bearbeitungszeit von jeweils 60 Minuten. Die Schüler*innen bearbeiteten dabei Testhefte unterschiedlicher Schwierigkeitsgrade aus zwei der vier Kompetenzbereichen (IPK1-4 gemäß Tabelle 14.1; s. außerdem Abschnitt 14.5.1 *Testdesign*) sowie einen Schüler*innenfragebogen, für den eine Bearbeitungszeit von etwa 50 Minuten vorgesehen war. Schüler*innen, für die keine erforderliche Einwilligung zur Bearbeitung des Schüler*innenfragebogens vorlag, bearbeiteten stattdessen eine Stillarbeit. Als nationale Ergänzung wurde in Deutschland ein dritter Testabschnitt mit einer Bearbeitungszeit von etwa 35 Minuten durchgeführt (ZIB-Modul). Dieser umfasste Aufgaben zum Problemlösen, kreativen Denken sowie zu kognitiven Grundfähigkeiten (ZMK1-3 gemäß Tabelle 14.1).

An Förderschulen erfolgte die Durchführung grundsätzlich nach demselben Verfahren, jedoch mit verkürzten Testzeiten. Der zusätzliche Testabschnitt entfiel ebenso wie die Durchführung der internationalen Zusatzdomäne Englisch. Einschließlich der Einführungs- und Instruktionsphasen betrug die Dauer einer Testsitzung an allen Schulen außer an Förderschulen etwa 4 Stunden und 45 Minuten, an Förderschulen rund 2 Stunden und 30 Minuten. Für die Schüler*innen der FLA-Testgruppe war darüber hinaus ein zweiter Testtag mit einer Dauer von etwa 4 Stunden vorgesehen (ZMK4a oder b und ZMF1 gemäß Tabelle 14.1).

Abbildung 14.6: Schematische Darstellung zu Ablauf, Dauer und Inhalt der Testsitzungen an allgemeinen und berufsbildenden Schulen getrennt nach PISA-Kernstichprobe, PISA-FLA-Stichprobe und Stichprobe der Neuntklässler*innen (nationale Option)



* HV/SP = Hörverstehen und Sprechen

*Verkürzter Testtag an Förderschulen: Kompetenztests Teil 1 und Teil 2: jeweils 30 Min., Schüler*innenfragebogen: 20 Min., ZIB-Modul entfällt, Gesamtdauer 2 Stunden 30 Minuten*

Zur Sicherstellung eines standardisierten Ablaufs folgten die Testleiter*innen den Vorgaben des Testleitungsskripts, lasen Anweisungen wortwörtlich vor und achteten auf die Einhaltung der festgelegten Bearbeitungszeiten. Während der Testsitzung dokumentierten sie den Teilnahmestatus der Schüler*innen sowie besondere Vorkommnisse in standardisierten Listen und Protokollen. Nach Abschluss der Erhebung kontrollierten sie die Vollständigkeit der Erhebungsinstrumente und schickten diese an die IEA Hamburg zurück.

Nach der Durchführung aller Erhebungen an einer Schule wurde geprüft, ob die international vorgegebenen Anforderungen an die Teilnahmequote erfüllt waren. War dies nicht der Fall, wurde für die bei den regulären Testsitzungen abwesenden Schüler*innen ein testgruppenübergreifender Nachtest durchgeführt. Dieser folgte denselben Abläufen und Standards wie die regulären Testsitzungen.

Exemplarischer Ablauf eines PISA-Testtages Testgruppe 4 (Zusatzdomäne FLA)

Die Testgruppe der Zusatzdomäne FLA (Testgruppe 4) wies im Vergleich zu den anderen Testgruppen einige besondere organisatorische und inhaltliche Merkmale auf, die im Folgenden skizziert werden (Abbildung 14.6).

Die Erhebung erstreckte sich über zwei aufeinanderfolgende Tage. Am ersten Testtag bearbeiteten die Schüler*innen Leseaufgaben in Deutsch und Englisch sowie Aufgaben zum Hörverstehen und Sprechen in Englisch (IOK1 gemäß Tabelle 14.1). Daran schlossen sich der Schüler*innenfragebogen sowie eine Schreibaufgabe auf Englisch im Rahmen der Begleitforschungsstudie PISA-Schreiben an (ZMK4a oder b gemäß Tabelle 14.1). Die Gesamtdauer der Erhebung entsprach derjenigen den anderen Testgruppen.

Am zweiten Testtag bearbeiteten die Schüler*innen auf Deutsch Kompetenzaufgaben aus den Bereichen Mathematik, Lesen und Naturwissenschaften, die den PISA-Kerndomänen entsprachen. Anschließend folgte, wie bereits am ersten Testtag, eine Schreibaufgabe auf Englisch (s. Abschnitt 14.2). Die Besonderheit dieser Schreibaufgaben bestand darin, dass sie an einem der beiden Tage mit Einsatz von künstlicher Intelligenz und am anderen Tag ohne Einsatz von künstlicher Intelligenz bearbeitet wurden. Der zweite Testtag war mit einer Gesamtdauer von rund 4 Stunden etwas kürzer als der Erste, da hier die Bearbeitung des Schüler*innenfragebogens entfiel.

Zusammengefasst unterschied sich der Ablauf der PISA-FLA-Testgruppe von einem regulären Testtag insbesondere durch die Verteilung der Testung auf zwei Tage, die reduzierte Gruppengröße von maximal elf Schüler*innen, den zusätzlichen Einsatz von Headsets sowie die spezifischen Aufgabenformate und die daraus resultierenden organisatorischen Maßnahmen.

14.5 Testdesign, Skalierung und Messung

Philipp Sterner

Im folgenden Abschnitt werden die Grundlagen des Testdesigns sowie die Verarbeitung der aus der Testung gewonnenen Daten zu Kompetenzmessungen auf Schüler*innen-Ebene beschrieben. Die Erstellung des Testdesigns, die Skalierung der Testantworten und alle damit verbundenen Entscheidungen lagen in der Verantwortung des internationalen Auftragnehmers ACER. Entsprechend werden hier vorrangig die Grundideen hinter den PISA-Messungen erläutert; Details finden sich in den Technical Reports der OECD (OECD, 2026b).

14.5.1 Testdesign

Als Testdesign bezeichnet man das Kombinieren einzelner Items (aus den Kompetenztests sowie Schüler*innen- bzw. Hintergrundfragebögen) zu einem größeren Test und die anschließende Zuordnung dieser Tests zu den Testpersonen, bei PISA also Schüler*innen. Wie in vielen groß angelegten ▶ Studien des Bildungsmonitorings ist auch bei PISA sowohl die Stichprobe der Schüler*innen als auch die Zuordnung der Items, die die Schüler*innen bearbeiten, wahrscheinlichkeitsbasiert. Das bedeutet, dass jede*r Schüler*in nur eine zufällige Auswahl an Items vorgelegt bekommt, wobei sich die Menge an bearbeiteten Items zwischen Schüler*innen überschneidet. Dieses Prinzip des Matrixsamplings (Rutkowski et al., 2014) ermöglicht die Testung eines breiten Kompetenzspektrums mit ausreichend vielen Items, während die individuelle Testzeit pro Schüler*in dennoch im Rahmen gehalten

wird. Allerdings führt es auch dazu, dass die Inferenzeinheit, also die Ebene, auf der Aussagen getroffen werden können, ausschließlich die teilnehmenden Staaten und Gruppen (z. B. Geschlechter) innerhalb der Staaten sind. Kompetenzmessungen der Schüler*innen können also nur aggregiert auf Gruppenebene interpretiert werden und dürfen nicht als individuelle Kompetenzmessung missverstanden werden.

Eine Besonderheit des Messens mit Matrixsampling sind die daraus resultierenden unvollständigen Testhefte. Mehrere Items werden innerhalb ihrer jeweiligen Domäne (z. B. Lesen oder Naturwissenschaften) zu Clustern zusammengefasst. Diese Cluster werden dann auf verschiedene Testhefte aufgeteilt und alle Schüler*innen durchlaufen sogenannte Testpfade für zwei Domänen. Ein grob vereinfachtes Beispiel wäre wie folgt: Schüler*in A bearbeitet Items aus den Domänen Mathematik und Lesen, Schüler*innen B und C bearbeiten jeweils Items aus den Domänen Mathematik und Naturwissenschaften und so weiter. Die Items aus Mathematik und Naturwissenschaften, die die Schüler*innen B und C bearbeiten, müssen hierbei nicht exakt identisch sein. Die Testungen finden am Computer statt, es handelt sich also nicht um ausgedruckte, sondern digitale Testhefte (Details s. Abschnitt 14.5.3 zu computerbasierter Testung). Am Ende der Testungen liegen für alle Schüler*innen Testantworten zu zwei Domänen sowie Angaben zu Fragen aus den Hintergrundfragebögen vor.

Aus diesen vorliegenden Testantworten und Informationen müssen Kompetenzwerte gewonnen werden, um anschließende Analysen wie Mittelwertvergleiche auf Ebene der teilnehmenden Staaten sowie Gruppen innerhalb dieser anstellen zu können. Eine entscheidende Herausforderung hierbei ist jedoch, dass die Antworten aus den Kompetenztests nicht ohne Weiteres aggregiert werden können, etwa durch Aufsummieren der richtigen Antworten (wie z. B. bei einer Schulaufgabe). Der Grund dafür ist, dass durch das Testdesign nicht alle Schüler*innen dieselben Items bearbeiten. Unterschiede in der Kompetenz könnten also auch auf unterschiedliche Schwierigkeitsgrade der Items zurückzuführen sein. Außerdem hätten dann nicht alle Schüler*innen Kompetenzwerte für alle Domänen, sondern nur für die Domänen, in denen sie auch Items bearbeitet haben.

Diese beiden Herausforderungen, also unterschiedliche Schwierigkeiten von Items sowie unvollständige Daten durch unvollständige Testhefte, werden durch eine testtheoretische Modellierung der Daten mit Methoden aus der Item-Response-Theorie (IRT; Cai et al., 2016; Thissen, 2025; Wilson, 2023) angegangen. Das Resultat dieser Modellierung sind Kompetenzwerte für alle Schüler*innen für alle Domänen auf einer für alle Staaten und PISA-Erhebungsrunden gemeinsamen Skala beziehungsweise Einheit.

14.5.2 Parameterschätzung und Skalierung im Sinne der IRT

Das Grundprinzip der IRT besteht in der Modellierung der Antwortwahrscheinlichkeiten von Personen auf Items. Das einfachste Beispiel ist ein dichotomes Item mit zwei Antwortmöglichkeiten: richtig ($x = 1$) und falsch ($x = 0$). Die Lösungswahrscheinlichkeit, also die Wahrscheinlichkeit, dass eine Person die richtige Antwort ankreuzt oder auswählt, hängt im einfachsten Fall von zwei Faktoren ab: der Schwierigkeit des Items und der Fähigkeit[5] (oder Kompetenz) der getesteten Person. Somit wird die Itemantwort zum „Wettkampf" zwischen Item und Person. Ist die Itemschwierigkeit höher als die Fähigkeit der Person, so sollte die Wahrscheinlichkeit, dass die Person richtig antwortet, kleiner sein als die Wahrscheinlichkeit, dass sie falsch antwortet; für eine höhere Fähigkeit der Person als die Schwierigkeit des Items sollte es umgekehrt sein. Es lässt sich zudem zeigen, dass Items auf Basis dieser Größen unterschiedliche Informationen beziehungsweise Erkenntnisgewinne über die Kom-

5 In diesem Abschnitt wird einheitlich der Begriff „Fähigkeit" verwendet, da dieser in der psychometrischen Literatur zur IRT etabliert ist. Er bezieht sich hier jedoch auf die Kompetenzen der Schüler*innen.

petenz einer Person zulassen. Items ermöglichen den optimalen Informationsgewinn, wenn ihre Schwierigkeit genau der Fähigkeit der Person entspricht. In diesem Fall beträgt die Lösungswahrscheinlichkeit genau 50 Prozent.

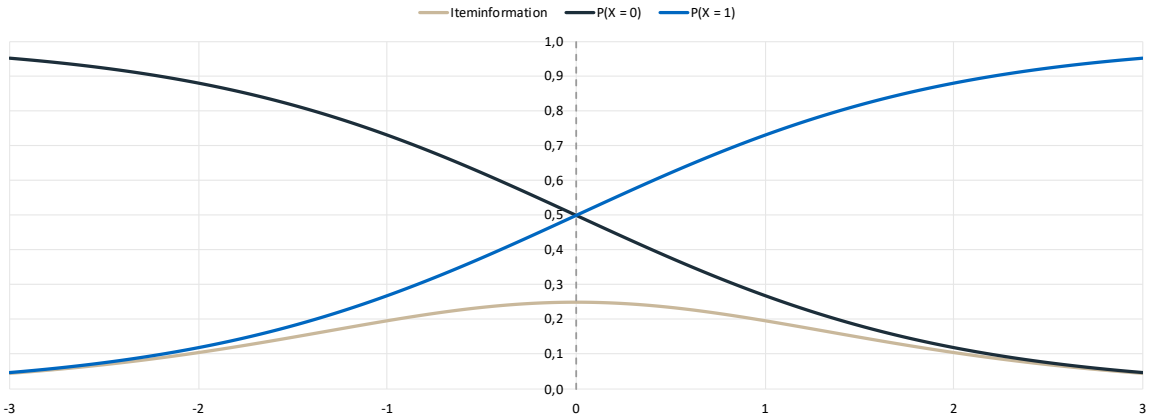
Um die Itemschwierigkeit und die Personenfähigkeit direkt miteinander vergleichen zu können, müssen sie allerdings eine gemeinsame Skala aufweisen. Dies wird in der IRT durch Skalierungsmodelle ermöglicht. Die Modellierung in PISA basiert auf dem ▶ 2-parameter-logistic Model (2-PL-Modell; Birnbaum, 1968). Dieses Modell hat auf Itemseite zwei ▶ Modellparameter: die oben erwähnte Itemschwierigkeit und die ▶ Itemdiskrimination. Die Itemdiskrimination hängt mit der Steigung der Modellfunktion (▶ Item Characteristic Curve, ICC) zusammen. Wie erwähnt, ist die Iteminformation maximal, wenn Schwierigkeit und Fähigkeit gleich groß sind; also mathematisch gesprochen am Wendepunkt der ICC. Zusätzlich bestimmt die Diskrimination, an welchem Punkt der kontinuierlichen Schwierigkeits-Fähigkeits-Skala größere oder kleinere Unterschiede in den Lösungswahrscheinlichkeiten zwischen Personen mit unterschiedlichen Fähigkeitswerten zu erwarten sind. Die Antwortwahrscheinlichkeiten einer Person *i* mit der Fähigkeit θ_i auf einem Item *j* mit der Schwierigkeit σ_j und Diskrimination α_j (mit $\alpha_j \in]0; +\infty[$) sind somit

$$P(X_{ij} = x_{ij} | \theta_i, \sigma_j, \alpha_j) = \frac{e^{x_{ij} \cdot \alpha_j(\theta_i - \sigma_j)}}{1 + e^{\alpha_j(\theta_i - \sigma_j)}},$$

wobei $x_{ij} \in \{0,1\}$. Hierbei zeigt sich, dass bei einem Item mit sehr hoher Diskrimination um den Wendepunkt herum (also um die Stelle σ_j) bereits bei kleinen Unterschieden zwischen den Fähigkeiten zweier Personen größere Unterschiede zwischen ihren Lösungswahrscheinlichkeiten erwartet werden. Für polytome Items, also Items mit mehr als zwei Antwortmöglichkeiten, werden bei PISA entsprechende Erweiterungen des Generalisierten Partial Credit Models (GPCM; Masters, 1982; Muraki, 1992) verwendet. Das GPCM ermöglicht die Modellierung von Antwortwahrscheinlichkeiten bei Aufgaben mit mehr als zwei Antwortmöglichkeiten. Diese Art von Items ist bei den Kompetenzmessungen in PISA der Regelfall, sodass bei vielen Items auch Punkte für Teillösungen vergeben werden können.

Graphisch lassen sich diese Prinzipien durch die ICC sowie die Funktion der Iteminformation veranschaulichen. Abbildung 14.7 zeigt die ICCs eines dichotomen Items mit $\alpha = 1$ sowie $\sigma = 0$ (auf einer Logit-Skala, siehe x-Achse der Grafik). Die y-Achse zeigt die Antwortwahrscheinlichkeiten für $x = 1$ und $x = 0$, sowie die Iteminformation. Hierbei wird ersichtlich, dass ein dichotomes Item dann den maximalen Informationswert liefert, wenn die Lösungswahrscheinlichkeit genau 0,5 beträgt; also wenn es einer Person vorgelegt wird, deren Fähigkeit genau der Itemschwierigkeit entspricht (hier auf der Logit-Skala also auch 0). Für die Messgenauigkeit von Tests bedeutet dies praktisch, dass man Schüler*innen idealerweise immer Items vorlegt, bei denen die Wahrscheinlichkeit, dass sie das Item lösen, genau 0,5 beträgt. Man sieht außerdem auch, dass der Informationsgehalt und damit die Messgenauigkeit eines Items abnimmt, je weiter eine Person in ihrer Fähigkeit von der Itemschwierigkeit entfernt ist. Vereinfacht gesagt bedeutet dies, dass man durch ein extrem leichtes Item nichts oder kaum etwas über die Fähigkeit von Schüler*innen mit extrem hoher Fähigkeit lernen kann; die Lösungswahrscheinlichkeiten lägen nämlich nahe bei 1 (und umgekehrt für schwere Items bei Schüler*innen mit niedriger Fähigkeit, wobei die Lösungswahrscheinlichkeit nahe 0 läge).

Abbildung 14.7: Item Characteristic Curves eines Items mit Schwierigkeit 0 (senkrechte gestrichelte Linie) und Diskrimination 1 sowie die entsprechende Iteminformation.



Selbstverständlich ist es nicht möglich, allen Schüler*innen stets ein optimales Item zu präsentieren. Darin zeigt sich auch die große Herausforderung in der Testkonstruktion bei Studien wie PISA. Da die Messgenauigkeit eines Tests von der Fähigkeit der getesteten Person abhängt und vor allem in den besonders hohen und niedrigen Kompetenzbereichen eher gering ist, besteht die Notwendigkeit eines diversen und breit angelegten Itempools. Dieser muss sich durch Items mit unterschiedlichen Schwierigkeiten über den gesamten Kompetenzbereich hinweg und mit unterschiedlichen Diskriminationen auszeichnen. Nur so kann gewährleistet werden, dass über alle Staaten hinweg sowohl Schüler*innen mit hohen als auch mit mittleren und niedrigen Kompetenzausprägungen reliabel getestet werden können. Im Folgenden werden das Prinzip der computerbasierten sowie der adaptiven Testung beschrieben, welche es ermöglichen, die Testheft- und Itemauswahl zu Teilen an die Fähigkeit der getesteten Person anzupassen.

14.5.3 Computerbasierte und adaptive Testung

Seit PISA 2015 wird die Haupterhebung in den meisten teilnehmenden Staaten, so auch in Deutschland, vollständig computerbasiert durchgeführt (nachdem bereits zuvor einzelne Komponenten der Testung computerbasiert erprobt wurden). Dies erleichtert nicht nur das Sammeln und Auswerten der Antworten und somit der Daten; es ermöglicht auch ein aus psychometrischer Sicht vorteilhaftes Testdesign (d. h. Anordnen und Darbieten der Items). Für die Kompetenzmessung bei PISA wird auf ein mehrstufiges adaptives Testdesign (Yamamoto et al., 2019) zurückgegriffen. Vereinfacht gesagt bedeutet adaptives Testen, dass einer Person sukzessive die Items vorgelegt werden, die basierend auf der aktuell gemessenen Fähigkeit der Person eine möglichst hohe Messgenauigkeit aufweisen und hinsichtlich der Schwierigkeit am ehesten ihrer Fähigkeit entsprechen. Dies steht im Kontrast zu nicht-adaptivem, also linearem, Testen, bei dem die Testpfade beziehungsweise Anordnung der Items im Vorhinein feststehen.

Bei PISA 2025 werden pro Domäne alle Items in drei nicht überlappende Sets aufgeteilt, die hinsichtlich der Schwierigkeit und Aufgabenart äquivalent sind. Aus diesen Sets werden die oben bereits erwähnten digitalen Testhefte erstellt. Im Rahmen des mehrstufigen adaptiven Testens durchlaufen alle Schüler*innen drei Stufen. In der ersten Stufe bearbeiten sie eines aus mehreren Kerntestheften. Basierend auf ihrem Abschneiden in diesem Testheft werden sie dann in der zweiten Stufe zu einem Testheft mit höherer oder niedri-

gerer Schwierigkeit weitergeleitet. Nach der Bearbeitung folgt dann die dritte Stufe, in der sie erneut basierend auf ihrem Abschneiden in der zweiten Stufe ein Testheft niedriger oder höherer Schwierigkeit (bei Lesen) beziehungsweise niedriger, mittlerer oder hoher Schwierigkeit (bei Mathematik und Naturwissenschaften) bearbeiten. Zusätzlich zu diesen adaptiven Testpfaden gibt es in den Domänen Mathematik und Naturwissenschaften auch lineare Testpfade; in der Domäne Lesen wurden alle Schüler*innen adaptiv getestet, wohingegen in der innovativen Domäne Lernen in der digitalen Welt alle Testungen linear waren. Die Bearbeitungsdauer der Testhefte in jeder Stufe beträgt 20 Minuten, sodass ein Testpfad eine Stunde dauert. Da jede*r Schüler*in in zwei Domänen getestet wird, dauert die Kompetenzmessung insgesamt zwei Stunden.

Die Zuweisung zu Testheften unterschiedlicher Schwierigkeitsgrade nach den einzelnen Stufen erfordert, dass unmittelbar ein Maß für die Fähigkeit der Schüler*innen berechnet werden kann. Dies hat zur Folge, dass nur Aufgaben, die auch automatisch (d. h. computerbasiert) bewertet werden können (z. B. klare richtig/falsch-Items), in das Routing zu den jeweiligen Testpfaden einfließen können. Jene Aufgaben, die eine Bewertung durch einen Menschen erfordern, werden anschließend bewertet und fließen nicht in die Zuordnung der Testhefte während der adaptiven Testung ein. Außerdem werden in den Stufen 2 und 3 nicht allen Schüler*innen Testhefte basierend auf ihrer Kompetenz zugewiesen. Einem kleinen Teil der Schüler*innen wird wahrscheinlichkeitsbasiert, also unabhängig von der tatsächlich gezeigten Fähigkeit, ein schwieriges oder leichtes Testheft zur Bearbeitung vorgelegt. Ziel dieses Vorgehens ist es, für alle Items eine ausreichende Anzahl an Antworten zu erhalten. Dies ist notwendig für die oben beschriebene Skalierung mittels IRT-Modelle. Würde die Itemauswahl ausschließlich auf der Passung aus Kompetenz und Schwierigkeit basieren, würden einige Items nur selten genutzt beziehungsweise präsentiert (für Details s. van der Linden & Glas, 2010). Details zum testheftbasierten adaptiven Testen bei PISA, vor allem zu den Besonderheiten in den einzelnen Domänen, finden sich im Technical Report (OECD, 2026b).

14.5.4 Plausible Values

Bisher wurden das Testdesign sowie die testtheoretischen Grundlagen der Skalierung der Testdaten beschrieben. Das eigentliche Ziel der Testung und der oben beschriebenen psychometrischen Modellierung ist es aber, Schätzwerte für die Kompetenz von Schüler*innen (d. h. für θ_i) zu erhalten. Der Vorteil testtheoretischer Skalierungsmodelle ist, dass sie die Mess(un)genauigkeit auf Personenebene berücksichtigen; im Übrigen ein weiterer Umstand, dem aufsummierte oder gemittelte „Punkte" (im Sinne richtiger Antworten) nicht Rechnung tragen würden. Bei PISA werden die Kompetenzen der Schüler*innen nach dem Prinzip der multiplen Imputation ermittelt; ein Vorgehen sehr ähnlich zur Imputation fehlender Werte (s. Enders, 2022). Konkret bedeutet dies, dass für alle Schüler*innen basierend auf A-posteriori-Verteilungen sogenannte Plausible Values (plausible Werte) ermittelt werden (Mislevy, 1991; Mislevy et al., 1992). Hierfür werden zunächst aus einer Kombination der Testantworten (modelliert durch IRT-Skalierungsmodelle) und der Informationen aus den Hintergrundfragebögen der Schüler*innen (modelliert durch ▶ latente Regressionen) für alle Schüler*innen individuelle A-posteriori-Verteilungen der Kompetenzdomänen geschätzt. Anschließend werden für alle Schüler*innen jeweils zehn Plausible Values pro Kompetenzdomäne aus diesen Verteilungen gezogen. Das Einbeziehen sowohl von Antworten aus den Kompetenztests als auch von Hintergrundfragebögen hat entscheidende Vorteile: Zum einen stellt es sicher, dass bei der Imputation der Kompetenzwerte die Verbindung zu interessierenden Hintergrundvariablen (wie z. B. Geschlecht oder soziale Herkunft) be-

rücksichtigt wird beziehungsweise erhalten bleibt. Das ermöglicht spätere Analysen der Zusammenhänge zwischen Kompetenzen und diesen Hintergrundvariablen, wie sie auch in diesem Berichtsband zu finden sind. Zum anderen ist es so möglich, Kompetenzschätzungen für alle Schüler*innen in allen Domänen zu ermitteln und gleichzeitig durch das Matrixsampling-Testdesign die Testzeit im Rahmen zu halten. Das heißt, dass Schüler*innen auch Kompetenzschätzungen für Domänen erhalten, in denen sie gar keine Testhefte bearbeitet haben. In diesen Fällen dominieren bei der Generierung der Verteilung der Kompetenzen die Informationen aus den Hintergrundfragebögen. Insgesamt lässt sich zeigen, dass diese Art der Kompetenzschätzung bei groß angelegten Studien wie PISA unverzerrte Analysen wie Mittelwertvergleiche zwischen Staaten ermöglicht, die den Messfehler auf Personenebene berücksichtigen. Diese Schätzungen dürfen jedoch nicht als individuelle Punktschätzer der Kompetenzen von Personen missverstanden werden. Sie eignen sich nicht zur individuellen Diagnostik. Unverzerrte Inferenz ist nur auf Ebene von Gruppen (z. B. Staaten oder Geschlecht innerhalb der Staaten) möglich.

Für statistische Analysen mit Plausible Values ergibt sich aus dem Prinzip der multiplen Imputation außerdem eine Besonderheit: Jede Analyse muss zehnmal wiederholt werden, da sie jeweils pro Plausible Value gerechnet wird, und die Ergebnisse dieser Berechnungen (z. B. Mittelwerte oder Regressionsparameter) werden im Anschluss gemittelt (zum genauen Vorgehen s. Rubin, 1987, sowie OECD, 2009). Für die Punktschätzungen bedeutet dies, dass die zehn Schätzungen ungewichtet gemittelt werden. Für die ▶ Varianz beziehungsweise den Standardfehler dieser Schätzfunktionen müssen hingegen zwei Quellen statistischer Unsicherheit kombiniert werden, nämlich die Unsicherheit aufgrund der Stichprobenziehung (siehe Abschnitt 14.3.4 *Präzision der PISA-Daten*) und die Streuung der Schätzungen zwischen den einzelnen Plausible Values. Diese zweite Quelle von Varianz in den Schätzungen würde missachtet werden, wenn man die Plausible Values zuerst pro Schüler*in mittelt und dann auf Gruppenebene verrechnet. Folglich würde man die Varianz unterschätzen und jede darauffolgende Inferenz wäre zu optimistisch (d. h. Konfidenzintervalle wären zu schmal und Hypothesentests wären zu liberal).

14.5.5 Linking zwischen PISA-Erhebungsrunden und Analysen im Zeitverlauf

PISA ist keine längsschnittliche, sondern eine querschnittliche Studie und Vergleiche sind primär zwischen Gruppen innerhalb derselben Erhebungsrunde möglich. Dennoch besteht ein hohes Interesse an Analysen im Zeitverlauf. Diese gehen der Frage nach, ob sich die Kompetenzen der Schüler*innen eines Staates oder mehrerer Staaten (z. B. gemittelt über alle OECD-Staaten, also im OECD-Durchschnitt) über die Jahre verändert haben. Um diese Vergleiche zu ermöglichen, wird bei der Ermittlung der Plausible Values sichergestellt, dass zugrunde liegende Populationsmodelle (d. h. die A-posteriori-Verteilungen) über die Erhebungsrunden hinweg verlinkt sind (Kolen & Brennan, 2014). Aus testtheoretischer Sicht bedeutet dies vereinfacht, dass die Parameter der Skalierungsmodelle (z. B. die oben beschriebene Schwierigkeit und Diskrimination) von sogenannten Anker- oder Trenditems in jeder Erhebungsrunde gleichgesetzt werden. Trenditems sind Items, die in jeder Erhebungsrunde erneut verwendet werden. Für neu hinzufügte Items hingegen werden erhebungsrunden-spezifische Parameter geschätzt.

Das Resultat dieser Linking-Strategie sind Plausible Values, die nicht nur querschnittlich (also z. B. für alle Staaten), sondern auch längsschnittlich dieselbe Metrik aufweisen. Als Folge dieses Vorgehens repräsentiert ein Lesekompetenzwert von 500 in den Erhebungsrunden 2009 oder 2022 dieselbe Kompetenzausprägung wie in der Erhebungsrunde 2025.

Ebenso sind die Abstände zwischen Kompetenzwerten für die verschiedenen Erhebungen gleich. Ein Unterschied von einem Punkt ist, unabhängig von der Position auf der Skala, näherungsweise derselbe Abstand in allen PISA-Erhebungsrunden. Die PISA-Kompetenzskalen wurden in der jeweiligen Hauptdomäne des ersten Erhebungszyklus so definiert, dass die OECD-Staaten einen Mittelwert von 500 Punkten und eine Standardabweichung von 100 Punkten aufwiesen. Diese Metrik wird anschließend immer wieder auf spätere Erhebungszyklen übertragen.

Allerdings ist diese Übertragung nicht exakt, zum Beispiel weil die über Erhebungsrunden hinweg fixierten Parameter nicht exakt gleich sind. Dieser Umstand ist eine weitere Quelle statistischer Unsicherheit und muss bei der Schätzung von Standardfehlern von Mittelwertdifferenzen im Zeitverlauf berücksichtigt werden; zum Beispiel bei der Analyse, ob sich die mittlere Lesekompetenz in Deutschland im Jahr 2025 im Vergleich zu 2022 signifikant verändert hat. Aus diesem Grund werden bei der Skalierung der Kompetenztests Link Errors ermittelt (OECD, 2026a). Dieser Link Error wird additiv zur Varianz- beziehungsweise Standardfehlerschätzung der Differenz zweier Mittelwerte hinzugefügt. Im obigen Beispiel zur Lese-Kompetenz im Jahr 2025 im Vergleich zu 2022 wäre der Standardfehler der Mittelwertdifferenz also

$$SE(\bar{X}_{2025} - \bar{X}_{2022}) = \sqrt{SE(\bar{X}_{2025})^2 + SE(\bar{X}_{2022})^2 + LinkError^2}\,.$$

Hierbei werden die Standardfehler der Mittelwertschätzungen für 2025 und 2022 jeweils wie oben beschrieben aus der Kombination der Stichprobenvarianz und der Imputationsvarianz berechnet. Durch die Hinzunahme des Link Errors wird die zusätzliche Unsicherheit aufgrund des Linkings berücksichtigt, und die daraus abgeleitete Inferenz (z. B. Konfidenzintervalle und Hypothesentests) ist valide.

Der Link Error wird für jede Kombination aus Vergleichen von Erhebungsrunden und für jede Hauptdomäne einzeln ermittelt und gilt anschließend für alle teilnehmenden Staaten. Das heißt beispielsweise, dass für den Vergleich der Erhebungsrunden 2018 und 2025 für die Domäne Lesen ein anderer Link Error gilt als für den Vergleich von 2022 und 2025 für die Domäne Lesen. Auch zwischen den Domänen unterscheiden sich die Link Errors. Allgemein bildet der Link Error potenzielle Verschiebungen auf den Kompetenzskalen ab, nachdem eine Kompetenz zum ersten Mal von einer Hauptdomäne zu einer Nebendomäne wurde. Folglich gibt es Link Error und damit die Möglichkeit zu Vergleichen im Zeitverlauf für Lesen bereits ab 2000, für Mathematik ab 2003 und für Naturwissenschaften erst ab 2006.

14.5.6 Modellierung von Hintergrundfragebögen

Wie oben beschrieben, werden die Daten aus den Hintergrundfragebögen (z. B. zu Geschlecht oder sozialer Herkunft) zunächst für die Ziehung der Plausible Values verwendet. Zusätzlich werden aus manchen dieser Antworten auf zusammenhängende Fragebogenitems im Anschluss Variablen abgeleitet. Diese Variablen dienen zum einen der Beschreibung der Stichprobe, etwa hinsichtlich des sozioökonomischen Hintergrunds. Zum anderen lassen sich auch Einstellungen und Merkmale über die Kompetenzen hinaus untersuchen, beispielsweise das schulische Zugehörigkeitsgefühl der Schüler*innen. Außerdem lassen sich Zusammenhänge zwischen diesen abgeleiteten Variablen und den Kompetenzen quantifizieren. Ein Beispiel ist die Analyse, ob mehr elterliche Unterstützung mit höherer Kompetenz der Schüler*innen einhergeht.

Die Ableitung von Variablen aus einzelnen Fragebogenitems ist eine Form des psychologischen Messens. Hierbei werden, wie bei der Kompetenzmessung auch, mehrere Items

eines Merkmals verwendet, um basierend auf einem testtheoretischen Modell individuelle Werte von Personen auf diesem Merkmal zu ermitteln. Anders als bei der Kompetenzmessung werden allerdings nicht mehrere Plausible Values pro Person gezogen, sondern es wird basierend auf einem Messmodell nur ein konkreter Wert pro Person geschätzt. Vereinfacht gesagt werden bei PISA zwei Arten von Konzeptualisierungen dieser Messmodelle betrachtet: formative und reflektive Messungen.

Formative Messungen (z. B. Bollen & Lennox, 1991) basieren auf einer einfachen Verrechnung mehrerer Items (oder gebildeter Variablen) zu einer weiteren Variable (oder Index). Das wohl prominenteste Beispiel bei PISA ist der sogenannte Economic, Social and Cultural Status (ESCS), also der sozioökonomische Hintergrund. Der ESCS ist eine standardisierte Zusammenfassung aus drei anderen Variablen zu häuslichen Besitzgütern (HOMEPOS) sowie zur Bildung (PAREDINT) und zum Beruf der Eltern (HISEI). Aus testtheoretischer Sicht ist bei formativer Messung zu beachten, dass sich die Merkmalsdefinition (hier also die Definition des ESCS) aus den Variablen ergibt, mit denen das Merkmal gemessen wird. Würde man für die Messung andere Variablen verwenden, so hätte die abgeleitete Variable ESCS eine andere Bedeutung. Da sich die Bestandteile des ESCS über die PISA-Erhebungsrunden hinweg tatsächlich mehrfach geändert haben, sind direkte quantitative Vergleiche über die Zeit in diesem Fall nicht sinnvoll.

Reflektive Messmodelle (Lord & Novick, 1968) werden insbesondere für psychologische Merkmale angenommen. Der Grundgedanke ist, dass ein Merkmal ursächlich die Antworten auf Items beeinflusst und dass dieser Einfluss durch testtheoretische Modellierung quantifiziert werden kann (Borsboom et al., 2004). Die oben beschriebene Kompetenzmessung und Skalierung durch IRT-Modelle sind Beispiele einer reflektiven Messung: Die Kompetenz θ einer Person „verursacht" eine bestimmte Itemantwort. Im Gegensatz zu formativen Messungen ändert sich die Definition des gemessenen Merkmals nicht durch die Itemwahl. Vielmehr basiert die Messung auf der theoretischen Annahme, dass alle verwendeten Items eine zufällige Auswahl aus einem „unendlichen Itempool" darstellen. Nach der Schätzung der Itemparameter auf Basis der Daten können dann Werte für θ, also für das zu messende Merkmal, abgeleitet werden.

Für reflektive Messungen werden bei PISA zur Modellierung der Daten aus den Hintergrundfragebögen ebenfalls IRT-Modelle eingesetzt und aus diesen ▶ Weighted Likelihood Estimates (WLE; Warm, 1989) abgeleitet. WLEs sind, anders als Plausible Values, feste Werte für Personen auf einem Merkmal. Die WLEs werden nach ihrer Schätzung am OECD-Durchschnitt standardisiert und haben somit eine Einheit, die von der ursprünglichen Antwortskala der zur Messung verwendeten Items unabhängig ist. Einen Sonderfall hierbei stellen trendskalierte WLEs dar. Diese werden, ähnlich wie die Kompetenzen, aus IRT-Modellen mit fixierten Itemparametern gewonnen und am OECD-Durchschnitt eines früheren Erhebungsjahres standardisiert. Somit werden hier ebenfalls Analysen im Zeitverlauf ermöglicht, da die WLEs über die Erhebungsrunden hinweg eine gemeinsame Metrik aufweisen. In Analysen in diesem Berichtsband wurde entsprechend kenntlich gemacht, ob es sich um neu- oder um trendskalierte WLEs handelt. Ein WLE von 0 bedeutet folglich, dass der Wert der Person auf dem zu messenden Merkmal dem OECD-Durchschnitt des Jahres entspricht, an dem der WLE standardisiert wurde; ein Wert von 1 oder −1 bedeutet, dass der Wert eine Standardabweichung über beziehungsweise unter dem entsprechenden OECD-Durchschnitt liegt. Aufgrund fehlender Werte bei den OECD-Staaten können für einige WLEs leichte Abweichungen von der intendierten Metrik auftreten, sprich der OECD-Durchschnitt ist unter Umständen nicht immer exakt 0.

Ein klassisches Beispiel eines WLEs und damit einer reflektiven Messung eines Merkmals bei PISA ist das schulische Zugehörigkeitsgefühl (BELONG). Dieses basiert auf einer Mes-

sung durch sechs Items, die alle unterschiedliche, wenngleich ähnliche Aspekte des schulischen Zugehörigkeitsgefühls erfragen.

Abschließend zeigen die Erläuterungen in diesem Abschnitt, dass PISA hohe testtheoretische Standards bei der Messung und Schätzung von Kompetenzen umsetzt und diese zum Teil auch definiert. Durch die Kombination aus Skalierungen im Sinne der IRT und Informationen aus den Hintergrundfragebögen ist es möglich, Kompetenzwerte für alle Schüler*innen zu ermitteln und diese auf Populationsebene zu vergleichen. Durch das Einbeziehen der Hintergrundfragebögen in die Ermittlung der Kompetenzwerte ist es außerdem möglich, Zusammenhänge zwischen Merkmalen wie Geschlecht, schulischem Zugehörigkeitsgefühl oder sozioökonomischem Hintergrund und den Kompetenzen zu analysieren.

14.6 Umgang mit fehlenden Werten

Philipp Sterner & Oliver Lüdtke

Der internationale Auftragnehmer ACER stellte den teilnehmenden Staaten die internationale Datenbasis erstmals im März 2026 zur Verfügung. Diese Bereitstellung enthält unter anderem sowohl Plausible Values für die Kompetenzdomänen als auch die Daten der Hintergrundfragebögen. Hierbei werden die Einzelitems sowie die daraus abgeleiteten Indizes mitgeliefert (s. Abschnitt 14.5.6). Ein entscheidender Unterschied zwischen den Plausible Values und den Daten aus den Hintergrundfragebögen besteht darin, dass erstere stets vollständig sind, während bei letzteren fehlende Werte auftreten können. Fehlende Werte in den Hintergrundfragebögen und daraus abgeleiteten Indizes können vielfältige Ursachen haben. Zum Beispiel kann es aufgrund des Testdesigns vorkommen, dass ein Item einer Schülerin gar nicht präsentiert wurde. Ebenso könnte diese Schülerin auf ein vorgelegtes Item keine Antwort gegeben haben, etwa aufgrund von Unaufmerksamkeit oder mangelnder Motivation.

Der Umgang mit fehlenden Werten ist methodisch herausfordernd, zugleich aber enorm wichtig. Die Validität der aus den Daten geschlossenen Inferenz hängt unter Umständen stark davon ab, ob und wie fehlende Werte berücksichtigt wurden. Ein weit verbreitetes Rahmenmodell zu fehlenden Werten stammt von Rubin (1987), wonach fehlende Werte bezüglich des angenommenen Ausfallprozesses in drei Kategorien eingeteilt werden können (Enders, 2022):

- komplett zufälliges Fehlen (missing completely at random; MCAR): Das Fehlen einer Variable ist unabhängig von ihrer eigenen tatsächlichen Ausprägung und unabhängig von allen anderen beobachteten Variablen.
- zufälliges Fehlen (missing at random; MAR): Das Fehlen einer Variable kann von anderen beobachteten Variablen abhängen, ist aber unter Konstanthaltung dieser beobachteten Variablen unabhängig von der eigenen tatsächlichen Ausprägung.
- nicht-zufälliges Fehlen (missing not at random; MNAR): Das Fehlen einer Variable hängt auch dann noch von der eigenen tatsächlichen Ausprägung ab, selbst wenn alle beobachteten Variablen berücksichtigt werden.

Am Beispiel eines Items zum schulischen Zugehörigkeitsgefühl (BELONG) wäre komplett zufällig fehlend (MCAR) also dann gegeben, wenn ein Item aufgrund technischer Probleme nicht präsentiert wird. Die fehlende Antwort hängt dann weder vom tatsächlichen Zugehörigkeitsgefühl noch von anderen Variablen wie dem Geschlecht oder der Kompetenz ab. Zufälliges Fehlen (MAR) in diesem Beispiel bedeutet, dass Schüler*innen mit niedrigerer Kompetenz oder mit Zuwanderungshintergrund das Item häufiger unbeantwortet lassen. Berücksichtigt man aber diese und andere beobachtete Variablen, so hängt das Fehlen nicht

zusätzlich vom tatsächlichen schulischen Zugehörigkeitsgefühl ab. Nicht-zufälliges Fehlen (MNAR) läge vor, wenn Schüler*innen mit beispielsweise tatsächlich sehr niedrigem schulischem Zugehörigkeitsgefühl das Item häufiger auslassen. Selbst bei Kontrolle aller anderen beobachteten Werte, zum Beispiel Zuwanderungshintergrund und Lesekompetenz, würde das Fehlen noch von den unbeobachteten Werten des schulischen Zugehörigkeitsgefühls abhängen.

Welche dieser drei Ursachen genau auf einen fehlenden Wert zutrifft, lässt sich nur selten eindeutig bestimmen. Vielmehr ist es eine inhaltlich-theoretische Annahme, die zum Beispiel auf Wissen über das Testdesign oder auf Annahmen über die Teilnehmenden beruht. Entscheidend ist jedoch, dass bestimmte Möglichkeiten des Umgangs mit fehlenden Werten nur bei bestimmten Ursachen zu unverzerrter Inferenz führen. Der einfachste Umgang mit fehlenden Werten ist, Personen mit fehlenden Werten aus den Analysen auszuschließen. Wenn eine Person beispielsweise keinen Wert für ihr schulisches Zugehörigkeitsgefühl hat, würde diese Person und alle ihre Werte vollständig aus Analysen mit dieser Variable ausgeschlossen werden. Dieses Vorgehen liefert in den meisten Fällen nur dann unverzerrte Schätzungen, wenn die Daten komplett zufällig (MCAR) fehlen, was allerdings die strengste Annahme ist. Ein Vorgehen, das bereits mit der Annahme eines zufälligen Fehlens (MAR) auskommt und nicht komplett zufälliges Fehlen voraussetzt, ist multiple Imputation (vgl. Abschnitt 14.5.4 Plausible Values). Das Prinzip der multiplen Imputation sieht vor, fehlende Werte durch möglichst plausible Werte zu ersetzen, wobei vor allem die Zusammenhänge zu anderen Variablen berücksichtigt werden. Das ist dasselbe Prinzip, nach dem in PISA und anderen Studien des Bildungsmonitorings Kompetenzschätzungen ermittelt werden (weshalb diese auch Plausible Values heißen). Üblicherweise werden für einen fehlenden Wert mehrere plausible Werte ermittelt (daher der Name multiple Imputation), um die Unsicherheit dieses Vorgehens im Standardfehler von Schätzfunktionen zu berücksichtigen. Konkret setzt sich der Standardfehler von Schätzfunktionen bei imputierten Variablen nämlich aus der Unsicherheit durch die Stichprobenziehung und aus der Varianz zwischen den imputierten Werten zusammen. Diese zweite Quelle statistischer Unsicherheit könnte nicht quantifiziert werden, würde man nur einen einzigen Wert imputieren (d. h. single imputation).

Multiple Imputation der Schüler*innendaten

Im Folgenden werden die Grundlagen der Imputationen der Schüler*innendaten bei PISA 2025 in Deutschland beschrieben, bevor im nächsten Abschnitt auf die konkrete technische Umsetzung eingegangen wird. Für die Analysen im Rahmen dieses nationalen Berichtsbandes (konkret bei der Analyse der Daten der Schüler*innen aus Deutschland) wurde in diesem Berichtsband entschieden, fehlende Werte auf Ebene der Indizes und WLEs (s. Abschnitt 14.5.6 *Modellierung von Hintergrundfragebögen*) zu imputieren. WLEs werden aus zusammengehörenden Items einer Skala gebildet und sind bereits in der Datenlieferung enthalten. Es müssen allerdings für mindestens drei Items einer Skala Werte vorliegen, damit der entsprechende WLE gebildet wird. Wenn also nur für zwei oder weniger Items Werte vorliegen (und die restlichen Items fehlen), wird der WLE als fehlender Wert ausgewiesen und entsprechend nach dem hier beschriebenen Verfahren ZIB-seitig imputiert. Das heißt, dass die internationale Datenlieferung um imputierte Werte für diese Variablen angereichert wurde, wobei die Imputation nur auf die Daten der Schüler*innen in Deutschland bezogen ist. Für andere Staaten wurden keine fehlenden Werte imputiert. Für Vergleiche zwischen Staaten bedeutet das, dass hierfür auf die vom internationalen Auftragnehmer ACER bereitgestellte Datengrundlage zurückgegriffen wurde.[6]

6 Ergänzend sei erwähnt, dass in dieser internationalen Datenbasis bereits teilweise Imputationen von ACER vorgenommen wurden, zum Beispiel bei der Ermittlung des Economic, Social, and Cultural Status (ESCS). Weitere Details hierzu finden sich im Technical Report (OECD, 2026b).

Dieses Vorgehen des Imputierens auf Indexebene entspricht einem *transform, then impute*-Ansatz (van Buuren, 2018), der bewusst vor dem Hintergrund der einfacheren Handhabbarkeit getroffen wurde. Der Ansatz *impute, then transform*, also eine Imputation der Einzelitems und anschließende Reskalierung der WLEs, würde erfordern, die Parameter des ursprünglichen Skalierungsmodells zu kennen und dies für mehrere Staaten zu tun, um die WLEs wieder am OECD-Durchschnitt standardisieren zu können.

Eine wichtige Spezifikation bei multiplen Imputationen ist die des Imputationsmodells. Vereinfacht gesagt, wird dieses Modell anhand der vorhandenen Daten geschätzt und anschließend genutzt, um plausible Werte als Ersatz für die fehlenden Werte vorherzusagen. Das Imputationsmodell sollte so komplex spezifiziert sein, dass es die Zusammenhänge zwischen den interessierenden Variablen abbildet und diese Zusammenhänge somit in den imputierten Werten aufrechterhält. In diesem Fall ermöglicht das Prinzip der multiplen Imputation bereits unter zufälligem Fehlen (MAR) unverzerrte Inferenz. Wie bereits erwähnt, wird dieser Vorgang mehrfach wiederholt. Da die Kompetenzschätzungen bereits selbst multiple Imputationen der „fehlenden" Kompetenzwerte der Schüler*innen sind (alle Schüler*innen erhalten pro Domäne zehn Plausible Values), ergibt sich daraus ein verschachteltes Imputationsproblem: das Imputationsmodell muss einmal für jedes Set an Plausible Values und weiteren wichtigen Variablen geschätzt werden (also für die ersten Plausible Values der Domänen, dann für die zweiten usw.). Pro Imputationsmodell werden dann die multiplen Imputationen der fehlenden Werte auf der interessierenden Hintergrundvariable gezogen. Dies führt dazu, dass die multiplen Imputationen der Hintergrundvariable in den multiplen Imputationen der Kompetenzen (d. h. der Plausible Values) geschachtelt sind. Zwar gibt es Softwarelösungen, die damit umgehen können (z. B. das Paket BIFIEsurvey für das Statistikprogramm R; Robitzsch & Oberwimmer, 2026). Allerdings erhöht dies auch die Komplexität aller nachfolgenden Analysen. Aus diesem Grund wurde für die Imputation auch hier bewusst ein pragmatischer Ansatz gewählt: Für jedes der zehn Sets an Plausible Values wurde jeweils immer nur eine Imputation gezogen und dem finalen Datensatz hinzugefügt. Somit gab es für alle imputierten Hintergrundvariablen zehn multiple Imputationen auf derselben Ebene wie die zehn Plausible Values der Kompetenzen. Diese können dann nach der ohnehin für die Plausible Values verwendeten Rubin'schen Regel (Rubin, 1987) verrechnet und kombiniert werden (vgl. Abschnitt 14.5.4 Plausible Values). Der Vollständigkeit halber sei erwähnt, dass dieses Vorgehen die Varianz der imputierten Variablen unter Umständen unterschätzt, da die Varianz zwischen den Imputationen innerhalb eines Sets an Plausible Values nicht berücksichtigt werden kann. Dies dürfte sich also vor allem auf die Schätzung der Unsicherheit und weniger auf die Punktschätzungen auswirken. Der negative Effekt auf die Standardfehler von Schätzfunktionen (z. B. bei Regressionsparametern) könnte durch die große Stichprobe teilweise ausgeglichen werden.

Technische Umsetzung der multiplen Imputation und konkretes Imputationsmodell

Die technische Umsetzung der multiplen Imputation erfolgte in der Statistiksoftware R unter Verwendung der Pakete mice (van Buuren & Groothuis-Oudshoorn, 2011) und miceadds (Robitzsch & Grund, 2026). Ausgangspunkt waren die von ACER bereitgestellten Schüler*innendaten aus Deutschland. Zunächst wurden die für das Imputationsmodell relevanten Variablen ausgewählt. Hierzu zählten zum einen die Plausible Values der Domänen Naturwissenschaften, Mathematik, Lesen sowie zwei Subskalen der innovativen Domäne, also Computational Practices Prior Knowledge (CPPK; Vorwissenstest) und Computational Problem Solving (CMPS; deutsche Übersetzung: modellbasiertes Problemlösen). Zum anderen

wurden diejenigen Hintergrundvariablen einbezogen, die entweder selbst Ziel der Imputation waren oder als inhaltlich informative Hilfsvariablen zur Vorhersage fehlender Werte dienen konnten. Darüber hinaus wurden die Einzelitems derjenigen Indizes und WLEs aufgenommen, die in den Analysen verwendet werden sollten (d. h., die Ziel der Imputation waren). Auf diese Weise sollte sichergestellt werden, dass das Imputationsmodell möglichst viel der in den Daten enthaltenen Informationen zur Vorhersage fehlender Werte nutzt. Um Unterschiede zwischen Schulen (d. h. die Mehrebenenstruktur) zumindest näherungsweise zu berücksichtigen, wurde für jedes Plausible-Value-Set zusätzlich ein schulbezogener Mittelwert des jeweiligen Naturwissenschafts-Plausible-Values berechnet. Dieser Cluster-Mittelwert wurde dann als Hilfsvariable in das Imputationsmodell aufgenommen (für ausführlichere Möglichkeiten zur Berücksichtigung von Mehrebenenstrukturen s. Grund et al., 2018).

Zusätzlich wurden die finalen Schüler*innengewichte in das Imputationsmodell aufgenommen. Da Gewichte potenziell nichtlineare Zusammenhänge mit Antwortverhalten und Merkmalsausprägungen aufweisen können, wurden sie nicht als einzelne metrische Variable verwendet, sondern zunächst in Dezile eingeteilt und anschließend als Dummyvariablen kodiert. Diese Dummies gingen als weitere Hilfsvariablen in das Imputationsmodell ein. Die Gewichte selbst wurden dabei nicht mitaufgenommen.

Analog zur allgemeinen Imputationslogik wurde auch die technische Umsetzung nach Plausible-Value-Sets getrennt vorgenommen. Es wurden daher zehn getrennte Imputationsmodelle geschätzt, jeweils eines für das erste, zweite, dritte usw. Set an Plausible Values der Domänen Naturwissenschaften, Mathematik, Lesen, CPPK und CMPS. Alle übrigen Hintergrundvariablen blieben über diese zehn Modelle hinweg identisch. Auf diese Weise wurde gewährleistet, dass jede Imputation der Hintergrundvariablen konsistent auf genau einem Set an Kompetenzschätzungen beruhte.

Die eigentliche Imputation erfolgte mit der Funktion miceadds::mice.1chain(). Dabei handelt es sich um eine Variante des Fully-Conditional-Specification-Ansatzes, bei der nicht mehrere unabhängige Ketten, sondern eine lange Kette mit wiederholten Iterationen verwendet wird. Imputiert wurden alle zuvor festgelegten Zielvariablen sowie Variablen mit fehlenden Werten, die zwar nicht selbst Ziel der Imputation waren, aber als wichtige Hilfsvariablen in das Imputationsmodell mitaufgenommen wurden. Auf diese Weise wurde vermieden, dass unvollständige Hilfsvariablen die Stabilität der Imputationsmodelle beeinträchtigen.

Für die Imputation wurde durchgängig die Methode Partial Least Squares (PLS) verwendet (Grund et al., 2024). Diese dient der Dimensionsreduktion in Situationen, in denen eine große Zahl potenzieller Prädiktoren vorliegt und viele dieser Variablen hoch miteinander korrelieren. Der gesamte Informationsraum aus den Prädiktoren wird somit auf eine geringe Anzahl an Komponenten verdichtet. Im vorliegenden Fall wurden pro zu imputierender Variable 20 PLS-Komponenten verwendet. Die eigentliche Ziehung der fehlenden Werte erfolgte innerhalb dieses PLS-Rahmens mittels predictive mean matching (PMM). Dadurch wurde sichergestellt, dass imputierte Werte stets aus beobachteten, in den Daten tatsächlich vorkommenden Werten gezogen wurden.

Pro Set an Plausible Values wurde – entsprechend dem oben beschriebenen pragmatischen Vorgehen – nur eine Imputation[7] erzeugt. Der finale Datensatz enthielt somit für jede imputierte Hintergrundvariable zehn Versionen, die indexkonsistent zu den zehn Plausible Values der Kompetenzen vorliegen. Die nachfolgenden Analysen konnten damit in derselben Logik durchgeführt werden, die auch für Plausible Values generell verwendet wird.

7 Details zum Imputationsmodell sowie zu allen verwendeten Variablen können gerne per E-Mail (pisa@tum.de) angefragt werden.

Literatur

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S. 397–479). Addison-Wesley.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application, 3*(1), 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Enders, C. K. (2022). *Applied missing data analysis* (2. Aufl.). Guilford Press.

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology, 30*(4), 415. https://doi.org/10.2307/589632

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data at Level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics, 43*(3), 316–353. https://doi.org/10.3102/1076998617738087

Grund, S., Lüdtke, O., & Robitzsch, A. (2024). Multiple imputation of missing data in large studies with many variables: A fully conditional specification approach using partial least squares. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000694

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

Krauss, S., Lindl, A., Schilcher, A., Fricke, M., Göhring, A., Hofmann, B., Kirchhoff, P., & Mulder, R. H. (2017). *FALKO: Fachspezifische Lehrerkompetenzen*. Waxmann.

Kroehne, U. (2023). *Open computer-based assessment with the CBA ItemBuilder*. DIPF. https://doi.org/10.5281/ZENODO.10359757

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Hrsg.), *Reliability and validity of international large-scale assessment*. Springer. https://doi.org/10.1007/978-3-030-53081-5_7

Meinck, S., & Vandenplas, C. (2021). Sampling design in ILSA. In T. Nilsen, A. Stancel-Piątak, & J. E. Gustafsson (Hrsg.), *International handbook of comparative large-scale studies in education*. Springer. https://doi.org/10.1007/978-3-030-38298-8_25-1

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

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

OECD. (2009). *PISA data analysis manual: SPSS* (2. Aufl.). OECD Publishing. https://doi.org/10.1787/9789264056275-en

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. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematics, 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. (2021). *PISA 2025 foreign language assessment framework*. OECD Publishing.

OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2022-assessment-and-analytical-framework_dfe0b-f9c-en

OECD. (2024). *PISA 2022 technical report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

OECD. (2026a). *Technical standards PISA 2025*. https://www.oecd.org/pisa/pisaproducts/PISA_2025_Technical_Standards.pdf

OECD. (2026b). *Technical report PISA 2025*. OECD Publishing.

OECD. (2026c). *PISA 2025 assessment and analytical framework*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2025-assessment-and-analytical-framework_86c36975-en.html

Robitzsch, A., & Grund, S. (2026). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. https://CRAN.R-project.org/package=miceadds

Robitzsch, A., & Oberwimmer, K. (2026). *BIFIEsurvey: Tools for survey statistics in educational assessment*. https://CRAN.R-project.org/package=BIFIEsurvey

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). 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/b16061-9

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness*. Springer Netherlands. https://doi.org/10.1007/978-94-017-7459-8

Schiepe-Tiska, A., Heinle, A., Todtenhöfer, P., Reinhold, F., Neumann, K., & Reiss, K. (2026). Lern- und Leistungsaufgaben im mathematisch-naturwissenschaftlichen Unterricht: Eine Analyse der Orientierung an Bildungsstandards und des motivationalen Potenzials im Fach- und Schulartvergleich. *Zeitschrift für Erziehungswissenschaft*. https://doi.org/10.1007/s11618-026-01403-w

Thissen, D. (2025). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research, 60*(2), 211–235. https://doi.org/10.1080/00273171.2024.2396148

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

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

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

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2. Aufl.). Routledge. https://doi.org/10.4324/9781003286929

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

ZIB. (in Vorbereitung). *Skalenhandbuch PISA 2025*.

Zuehlke, O. (2011). Sampling design and implementation. In W. Schulz, J. Ainley, & J. Fraillon (Hrsg.), *ICCS 2009 technical report* (S. 59–68). International Association for the Evaluation of Educational Achievement (IEA).