Technische Hintergrundinformationen PISA 2018

Julia Mang, Sabrina Wagner, Jens Gomolka, André Schäfer, Sabine Meinck & Kristina Reiss

Das *Programme for International Student Assessment* (PISA) implementiert ein komplexes Stichproben- und Skalierungsdesign, das sich durch spezifische statistische Methoden bezüglich des Auswahlprozesses der Schulen und der Schülerinnen und Schüler sowie der Aufbereitung und Darstellung der erhobenen Antworten der PISA-Testung auszeichnet.

Die Stichprobenziehung der Schülerinnen und Schüler erfolgt in einem zweistufigen Verfahren. In einem ersten Schritt werden Schulen spezifisch je Bundesland und Schulart nach einem komplexen statistischen Zufallsverfahren ausgewählt. In einem zweiten Schritt werden innerhalb dieser Schulen ebenfalls per Zufallsstichprobe fünfzehnjährige Schülerinnen und Schüler zur Teilnahme an der PISA-Testung bestimmt. Weitere Qualitätsmerkmale der Stichprobenziehung werden eingehend erläutert. Hierzu zählen Angaben zu Teilnahmequoten sowie zur Repräsentativität der Stichprobe für die gesamte PISA-Zielgruppe in Deutschland.

Die Testantworten der Schülerinnen und Schüler werden in einem mehrstufigen Prozess, der die internationale Vergleichbarkeit der Ergebnisse sicherstellt, zur weiteren Verwendung aufbereitet. Hierfür eingesetzte Reliabilitätskriterien werden veranschaulicht. Ein statistisches Schätzverfahren wird eingesetzt, um die PISA-Kompetenzen auf der Grundlage der Schülerantworten zu berechnen. Die Ergebnisse werden anhand sogenannter Kompetenzstufen veranschaulicht.

Aussagen zur Vergleichbarkeit der unterschiedlichen PISA-Zyklen sowie Chancen und Herausforderungen kommender Erhebungen schließen dieses Kapitel ab.

1. Einleitung

In diesem Kapitel werden die technischen Grundlagen des *Programme for International Student Assessment* (PISA) für die Erhebung im Jahre 2018 eingehend erläutert.

Technisches beziehungsweise methodisches Ziel der Studie ist es, konsistente, präzise und generalisierbare Daten zu erhalten. Diverse Standards und Anforderungen an die Datenerhebung zur Erstellung einer internationalen Datenbank ermöglichen es, gültige Vergleiche und Schlussfolgerungen innerhalb und außerhalb Deutschlands zu ziehen (*Organisation for Economic Co-operation and Development* [OECD], 2017).

Neben grundlegenden Methoden, welche seit der ersten PISA-Erhebung im Jahr 2000 implementiert worden sind und somit auch die Vergleichbarkeit zwischen unterschiedlichen PISA-Erhebungen und -zyklen sicherstellen, werden auch Neuerungen dieser Befragung eingehend erläutert und im Ausblick auf zukünftige Erhebungen diskutiert.

Alle hier dargestellten Prozesse fokussieren auf Abläufe, welche in Deutschland durchgeführt werden. Für detaillierte Aktivitäten, welche vom internationalen Konsortium durchgeführt werden, sei auf den Technical Report der OECD verwiesen (OECD, 2017 sowie den Technical Report für PISA 2018).[1]

2. Stichprobenbeschreibung

In PISA 2018 wird wie in den vorherigen Erhebungszyklen eine systematische Teilerhebung realisiert. Hierfür erfolgt anhand exakter statistischer Regeln die Ziehung einer Stichprobe von fünfzehnjährigen Schülerinnen und Schülern, anhand derer nach Auswertung der Ergebnisse Verallgemeinerungen über die Grundgesamtheit, also alle fünfzehnjährigen Schülerinnen und Schüler in Deutschland, möglich sind (vgl. Bortz & Schuster, 2010; Brown, 2010; Häder, 2015; Kish, 1995; Levy & Lemeshaw, 2008; Thompson, 2012).

2.1. Populationsdefinitionen und Stichprobendesign

Um Rückschlüsse aus der stichprobenbasierten PISA-Erhebung auf die Grundgesamtheit der fünfzehnjährigen Schülerinnen und Schüler aller Teilnehmerstaaten zu ermöglichen, sowie die internationale Vergleichbarkeit zu sichern, sind Verfahren der Stichprobenziehung anzuwenden, die unverzerrte und präzise Populationsschätzer ermöglichen.

[1] Der Technical Report der OECD für PISA 2018 ist zur Publikation in 2020 vorgesehen.

In PISA werden in allen Teilnehmerstaaten zwei- oder mehrstufige Zufallsverfahren für die Ziehung der Stichprobe eingesetzt.[2] In der Regel werden in einem ersten Schritt Schulen gezogen und in einem zweiten Schritt Schülerinnen und Schüler in den teilnehmenden Schulen per Zufall ausgewählt. Dieses Verfahren wurde auch in Deutschland implementiert.

In Deutschland wurde das PISA 2018 Studiendesign noch durch eine zusätzliche Zielpopulation erweitert. Mit Hilfe dieser sollen auch Aussagen für *Schülerinnen und Schüler der 9. Klassenstufe* ermöglicht werden. Da sich beide Zielpopulationen, also Fünfzehnjährige und Neuntklässler, zumindest teilweise überlappen, wurde die Stichprobenziehung parallel für beide Populationen durchgeführt und soll im Folgenden detailliert beschrieben werden. Weiterhin wird auch die Stichprobenziehung für Lehrkräfte vorgestellt. Neben den Befragungen der Schülerinnen und Schüler sowie der Lehrkräfte wurden auch die Eltern der Jugendlichen und die Schulleiterinnen und Schulleiter der teilnehmenden Schulen befragt. Die Teilnehmer dieser Befragungen wurden direkt über die Schul- und Schülerziehung bestimmt, weshalb in diesem Bericht nicht weiter auf diese Zielgruppen eingegangen wird.

2.1.1. Schülerstichprobe

Wie auch in den vergangenen PISA-Erhebungen besteht die international vorgegebene Zielpopulation aus allen Schülerinnen und Schülern einer Alterskohorte. Bei dieser handelt es sich um alle Fünfzehnjährigen, die sich in der siebten oder einer höheren Klassenstufe befinden. Die genaue Definition der Altersgruppe fand in Abstimmung mit dem internationalen PISA-Konsortium statt und kann sich zwischen den Staaten aufgrund verschiedener Erhebungszeiträume leicht unterscheiden. Für Deutschland galt die folgende Definition der Zielpopulation für PISA 2018:

Teilnahmeberechtigt waren alle Schülerinnen und Schüler, die zwischen dem 1. Januar 2002 und dem 31. Dezember 2002 (einschließlich) geboren sind und die mindestens die 7. Klassenstufe oder eine höhere Klassenstufe besuchen.

Um vertiefende Analysen durchführen zu können, wurde diese Zielpopulation für Deutschland erweitert – und zwar um die 9. Klassen an allgemeinbildenden Schulen sowie Förderschulen.[3] Schülerinnen und Schüler in 9. Klassen können aus verschiedenen Altersgruppen stammen, also sowohl Fünfzehnjährige als auch Nichtfünfzehnjährige

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

[3] Berufsschulen waren nicht Teil der definierten Schulpopulation für diese Zusatzerhebung.

inkludieren. Somit konnten Fünfzehnjährige, die eine 9. Klasse besuchen, zumindest theoretisch Teil beider Stichproben sein: die der PISA-Basisstichprobe, sowie die der Erweiterungsstichprobe für 9. Klassen. Dieses Design ermöglicht vollumfängliche Repräsentativität für beide Zielpopulationen.

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

Somit setzte sich das Ziehungsverfahren für die Schülerstichprobe aus mehreren, teils parallel laufenden Schritten zusammen: Zunächst wurden die Schulen gezogen. Sie bildeten die primäre Stichprobeneinheit für beide Zielpopulationen auf Schülerebene. Anschließend erfolgte in einem zweiten Schritt die Ziehung einer vollständigen 9. Klasse an allgemeinbildenden Schulen. An Förderschulen wurden alle Neuntklässler zu einer einzelnen virtuellen 9. Klasse zusammengruppiert.[4] Innerhalb jeder teilnehmenden Schule wurden dann zunächst 30 Fünfzehnjährige gezogen. Diese Stichprobe bildete die Teilnehmerschaft für die reguläre PISA 2018 Stichprobe. Parallel zu diesem Ziehungsschritt wurden innerhalb der gezogenen Klasse an jeder Schule 15 Schülerinnen beziehungsweise Schüler per Zufall zusätzlich zu jenen Schülern ausgewählt, die bereits über die reguläre Schülerstichprobe gezogen worden waren. Hierbei konnte kein Jugendlicher doppelt gezogen werden. Das heißt, wurde ein Fünfzehnjähriger bereits für die Stichprobe der PISA-Population gezogen, konnte er nicht zusätzlich für die Erweiterungsstichprobe der Neuntklässler gezogen werden. Es konnten jedoch Fünfzehnjährige Teil der Klassenstichprobe sein, die *nicht* Teil der regulären PISA 2018 Stichprobe sind.[5]

Eine Übersicht über das Stichprobendesign bietet die folgende Abbildung 1:

[4] Da in Förderschulen generell weniger Neuntklässler anzutreffen sind sowie außerdem die übliche Klassenstruktur oft nicht gegeben ist, wurden hier alle Neuntklässler pro Förderschule zu einer gesamten 9. Klasse zusammengefügt.

[5] Die Daten dieser Schülerinnen und Schüler gehen in die Analysen beider Stichproben, also sowohl in die der PISA-Population als auch in die der Zusatzstichprobe, ein.

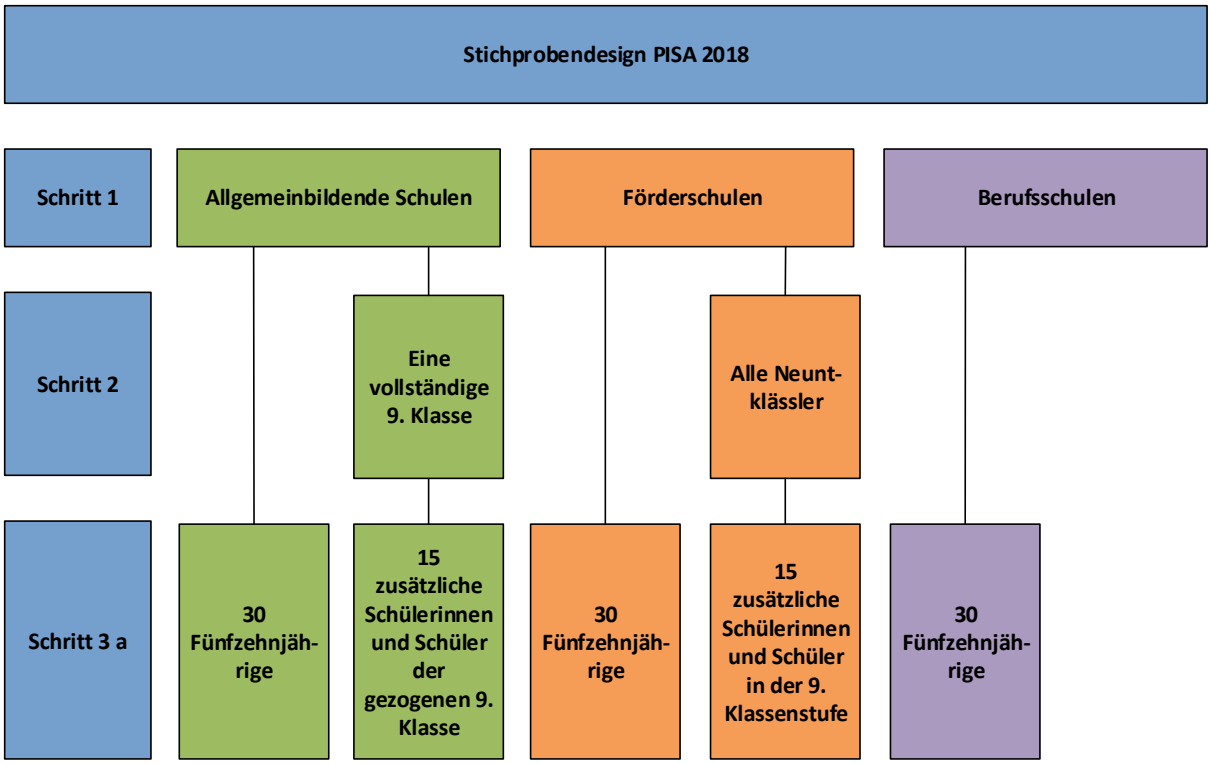

Abbildung 1: Stichprobendesign PISA 2018 – Schülerinnen und Schüler

2.1.2. Lehrerstichprobe

Deutschland nahm wie auch in der vergangenen Erhebung im Jahr 2015 an der zusätzlichen Lehrerbefragung teil.[6]

Wie auch bei der Stichprobe der Schülerinnen und Schüler wurde zur Ermittlung der Lehrerstichprobe die primäre Stichprobeneinheit, die Schulen, herangezogen. Es wurden also Lehrkräfte nur an den Schulen gezogen, welche für die PISA 2018 Testung ermittelt wurden.

Von besonderem Interesse sind in diesem Zusammenhang Lehrkräfte, welche Klassenstufen mit einem hohen Anteil an zur PISA-Grundgesamtheit (Fünfzehnjährige) gehörenden Schülerinnen und Schülern unterrichten. In Deutschland befinden sich insbesondere in den Klassenstufen 9 und 10 fünfzehnjährige Schülerinnen und Schüler (Statistisches Bundesamt, 2018). Die Definition der Zielpopulation der Lehrkräfte wurde daher wie folgt festgesetzt:

Die Lehrerbefragung ist an alle Lehrkräfte (inkl. Referendare) gerichtet, die eine Lehrbefähigung besitzen, um Schülerinnen und Schüler der 9. Jahrgangsstufe und/oder der

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

10. Jahrgangsstufe zu unterrichten – ungeachtet dessen, ob eine Lehrkraft dies aktuell tut, jemals getan hat oder künftig tun wird oder könnte. Vollzeit- sowie Teilzeitlehrkräfte, angestellte und verbeamtete Lehrkräfte sind dabei gleichermaßen zu berücksichtigen. Auch Lehrkräfte, die ihre Lehrtätigkeit an mehreren verschiedenen Schulen ausüben, sind für die Listung vorgesehen.
Als Lehrkraft gilt dabei eine Person, deren vorrangige oder hauptsächliche Aktivität in der Schule die Ausbildung von Schülerinnen und Schülern ist und die den Schülerinnen und Schülern Unterrichtsstunden erteilt. Lehrkräfte können mit den Schülerinnen und Schülern im ganzen Klassenverband in Klassenräumen arbeiten, in Kleingruppen oder im Einzelunterricht inner- oder außerhalb der regulären Klassenräume.

Entsprechend des internationalen Stichprobendesigns wurden zwei Gruppen von Lehrkräften unterschieden:

1) Lehrkräfte, welche Deutschunterricht geben (da Lesen 2018 die Hauptdomäne darstellt). In dieser Gruppe wurden 15 Personen pro Schule gezogen.
2) Lehrkräfte, welche sonstige Fächer unterrichten. In dieser Gruppe wurden 20 Personen pro Schule gezogen.

Um auch Analysen zum Klassenkontext der gezogenen 9. Klassen durchführen zu können, wurden Lehrkräfte an allgemeinbildenden Schulen, welche die jeweils gezogene Klasse im laufenden Schuljahr unterrichten, nach der Stichprobenziehung zusätzlich in die Lehrerstichprobe aufgenommen, auch wenn sie nicht Teil der Zufallsstichprobe waren.

Somit kann die bereits bekannte Darstellung des Stichprobendesigns für PISA 2018 wie folgt angepasst werden:

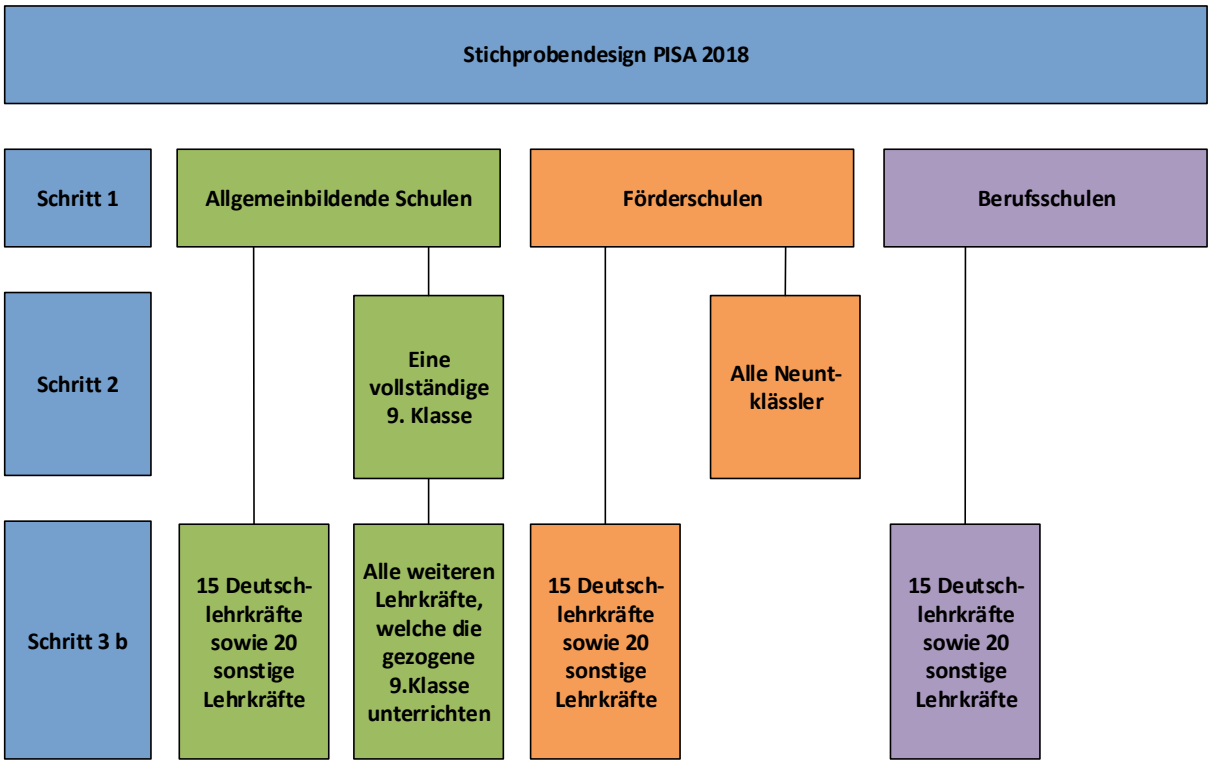

Abbildung 2: Stichprobendesign PISA 2018 – Lehrkräfte

2.2. Ablauf und Ergebnisse der Stichprobenziehungen

Im Folgenden Abschnitt wird erläutert, wie die Ziehungen der Schul-, Schüler- und Lehrerstichproben vorbereitet wurden. Außerdem werden die Ergebnisse der Ziehungen vorgestellt.

2.2.1. Ziehung der Schulstichprobe

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

Die Informationen zur Erstellung dieses *Sampling Frames* wurden bei den einzelnen statistischen Landesämtern eingeholt.[7] Die Listen umfassen alle Schultypen (allgemeinbildende Schulen, Förderschulen, Berufsschulen) im jeweiligen Land und enthalten die folgenden Angaben:

[7] Es gibt in Deutschland insgesamt 14 Statistische Landesämter, da für Berlin und Brandenburg sowie für Hamburg und Schleswig-Holstein jeweils ein Landesamt zuständig ist.

- die offizielle Schulnummer (die später im *Sampling Frame* pseudonymisiert wurde),
- die Schulart,
- die Anzahl der Schülerinnen und Schüler in den Geburtsjahrgängen 2001, 2002, 2003,
- die Anzahl der Schülerinnen und Schüler in den Klassenstufen 7 bis 10,
- die Anzahl der 7. bis 10. Klassen,
- die Trägerschaft (öffentlich oder privat),
- Informationen über Veränderungen der Schulart, Schulzusammenlegungen und Schulschließungen sowie
- für Förderschulen die Informationen über die Förderschwerpunkte gemäß den Vorgaben der Kultusministerkonferenz (KMK), wobei in Anlehnung an alle vorhergehenden PISA-Erhebungsrunden die Förderschwerpunkte Lernen, Sprache sowie emotionale und soziale Entwicklung berücksichtigt wurden.

Als Datengrundlage für die genannten Schulinformationen dienten die Schulstatistiken der einzelnen Länder für das Schuljahr 2016/2017. Falls diese für eine oder mehrere Informationen nicht verfügbar waren, wurde auf die jeweils zuletzt veröffentlichten Daten zurückgegriffen.

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

In Ermangelung eines exakten Wertes wurde dem *Sampling Frame* ein Schätzer der zu erwartenden Anzahl an Fünfzehnjährigen pro Schule im Jahr 2018 hinzugefügt, um die Durchführung des in PISA präferierten PPS-Ziehungsverfahrens möglich zu machen. Als Schätzer wurde die jeweilige Schüleranzahl des Geburtsjahrgangs 2001 verwendet, welche

außerdem mit den Schülerzahlen der Jahrgänge 2002 und 2003 abgeglichen wurde, um Schwankungen in den Geburtenzahlen zu identifizieren.[8]

Um Aussagen zum Stichprobenumfang hinsichtlich des erweiterten Forschungsdesigns zu ermöglichen, wurde im *Sampling Frame* außerdem die Anzahl zu erwartender Neuntklässler gelistet. Auch hierfür musste ein Schätzer verwendet werden, wobei es sich um die Anzahl der Neuntklässler an der jeweiligen Schule im Schuljahr 2016/2017 handelte. Um mögliche Inkonsistenzen ausgleichen zu können, wurden auch die Schülerzahlen der 8. und 10. Klassenstufe überprüft.

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

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

Eine Gruppe wird als Stratum bezeichnet und enthält einander „ähnliche“ Schulen, das heißt Schulen, die bestimmte Merkmale miteinander teilen. Die Stratifizierung kann explizit und implizit vorgenommen werden. Bei der expliziten Stratifizierung werden die Schulen in einzelne Gruppen beziehungsweise Strata aufgeteilt, welche unabhängig voneinander behandelt werden. Aus jedem Stratum wird dann eine separate Zufallsauswahl getroffen.

Bei der impliziten Stratifizierung werden die Schulen innerhalb eines expliziten Stratums im *Sampling Frame* noch einmal sortiert, um eine näherungsweise proportionale Verteilung der Schulen über sämtliche Strata innerhalb der Stichprobe zu gewährleisten.

In Deutschland kamen für PISA 2018 jeweils zwei explizite und implizite Stratifizierungsvariablen zum Einsatz:

Zunächst wurden alle Schulen, welche potenziell fünfzehnjährige Schülerinnen und Schüler unterrichten, in drei Strata aufgeteilt: allgemeinbildende Schulen, Förderschulen und

[8] Es wurde nicht der Geburtsjahrgang 2002 (PISA-Zielpopulation) verwendet, da so die Anzahl an Schülerinnen und Schülern, die sich 2018 an allgemeinbildenden Schulen befinden, überschätzt werden würde: Da in einigen Bundesländern viele Fünfzehnjährige die Hauptschule verlassen, um berufsbildende Schulen zu besuchen, dürfen nicht die Vierzehnjährigen als Schätzer für die Fünfzehnjährigen des nächsten Schuljahres verwendet werden.

Berufsschulen (1. explizite Stratifizierung). Anschließend wurden die allgemeinbildenden Schulen erneut in Gruppen aufgeteilt, und zwar den 16 Bundesländern entsprechend (2. explizite Stratifizierung). Damit ergeben sich insgesamt 18 explizite Strata. 16 davon repräsentieren allgemeinbildende Schulen in den einzelnen Bundesländern, ein Stratum enthält alle Förderschulen und eines enthält alle Berufsschulen. Die Förder- und Berufsschulen werden gesondert behandelt, da die relative Häufigkeit dieser Schulformen innerhalb der Bundesländer sehr unterschiedlich ist. Auch die im Verhältnis zu den anderen Schularten sehr geringe Anzahl an Schülerinnen und Schülern an Förder- und Berufsschulen erfordert die Berücksichtigung der beiden Schulformen als explizite Strata.

Weiterhin wurden die allgemeinbildenden Schulen in die Schulformen Hauptschule, Integrierte Gesamtschule, Realschule, Schule mit mehreren Bildungsgängen, Gymnasium und Schule-nicht-deutscher-Herkunftssprache eingeteilt (1. implizite Stratifizierung). Diese Maßnahme stellt sicher, dass sich die gezogenen allgemeinbildenden Schulen innerhalb der einzelnen Bundesländer näherungsweise proportional zur Gesamtzahl der Schülerinnen und Schüler auf die verschiedenen Schulformen verteilen. Um auch aus allen Bundesländern eine annähernd proportionale Menge zur tatsächlichen Anzahl an Schülerinnen und Schülern an Berufs- und Förderschulen in der Stichprobe zu haben, wurden diese beiden Strata außerdem jeweils nach Bundesländern geschichtet (2. implizite Stratifizierung).

Anhand dieser Vorgehensweise wurde sichergestellt, dass bei der anschließenden Ziehung der Schulstichprobe vom internationalen Konsortium eine Stichprobe gezogen wird, welche Schulen aus allen 18 Strata enthält und sich hinsichtlich der Anzahl an Schülerinnen und Schülern pro Bundesland und Schulform annähernd proportional zur Grundgesamtheit (Geburtenjahrgang 2002) verhält.

Insgesamt wurden 234 Schulen für PISA 2018 gezogen.

Tabelle 1 zeigt die Anzahl gezogener Schulen in den einzelnen Strata:

Tabelle 1:Bruttostichprobe der Schulen nach Bundesland und Schulart

Bundesland	HS	IG	MBG	R	G	ND	F	B	∑
Baden-Württemberg	8	4	0	11	11	0	3	4	***41***
Bayern	10	0	0	11	10	0	0	4	***35***
Berlin	0	5	0	0	3	0	1	0	***9***
Brandenburg	0	1	2	0	2	0	0	0	***5***
Bremen	0	1	0	0	1	0	0	0	***2***
Hamburg	0	3	0	0	1	0	0	0	***4***
Hessen	2	3	0	4	6	2	1	1	***19***
Mecklenburg-Vorpommern	0	0	2	0	2	0	1	0	***5***
Niedersachsen	2	3	4	5	8	0	1	1	***24***
Nordrhein-Westfalen	6	10	2	12	16	0	2	2	***50***
Rheinland-Pfalz	0	2	4	0	4	0	1	0	***11***
Saarland	0	1	0	0	1	0	0	1	***3***
Sachsen	0	0	4	0	4	0	1	0	***9***
Sachsen-Anhalt	0	0	2	0	2	0	0	0	***4***
Schleswig-Holstein	0	4	1	0	2	0	0	1	***8***
Thüringen	0	0	3	0	2	0	0	0	***5***
Gesamt	28	37	24	43	75	2	***11***	***14***	234

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

Wie die Tabelle zeigt, wird nicht aus jedem Stratum auch jede Schulform gezogen. Alle Schulen hatten jedoch eine von Null verschiedene Ziehungswahrscheinlichkeit.[9]

2.2.2. Ziehung der Schülerstichproben

Nach der Bestimmung der Schulen, an denen die Datenerhebung durchgeführt werden sollte, konnten dem oben vorgestellten Stichprobendesign entsprechend die Schüler-, Klassen- und Lehrerstichprobengezogen werden.

Um eine Schülerstichprobe zu erhalten, welche die Anforderungen des Studiendesigns erfüllt, mussten verschiedene demografische Daten herangezogen werden. Von den für PISA 2018 gezogenen Schulen wurden alle teilnahmeberechtigen Schülerinnen und Schüler, also sowohl alle Fünfzehnjährigen als auch alle Neuntklässler, gelistet. Diese Schülerlisten enthielten unter anderem die Merkmale Geburtsjahr, Geschlecht, Klassenstufe und Klassenbezeichnung. Personenbezogene Daten wie beispielsweise die Namen der Schüler verblieben dabei in den Schulen, um volle Pseudonymität sicherzustellen.

[9] Ausführliche Beispiele zur Schulstichprobenziehung sowie den dazugehörigen Ziehungsalgorithmen können den PISA Technical Reports unter: http://www.oecd.org/pisa/pisaproducts/ entnommen werden.

Aus diesen Angaben konnten außerdem Auflistungen aller 9. Klassen pro Schule bereitgestellt werden.

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

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

Für die Stichprobenziehung von Klassen und Schülern kam die vom internationalen Konsortium bereitgestellte Software KeyQuest in Anwendung.

Zunächst wurden die Klassenlisten in die Software eingelesen und die Klassenziehung durchgeführt. Für Förderschulen wurden wie bereits erwähnt alle Neuntklässler in einer virtuellen Klasse zusammengefasst.

Nachdem so je Schule eine 9. Klasse gezogen worden war, wurden auch die Schülerlisten in KeyQuest eingelesen und die Stichprobe der 30 Fünfzehnjährigen pro Schule gezogen. Anschließend wurden an allgemeinbildenden- und Förderschulen zusätzlich jeweils 15 Neuntklässler innerhalb der zuvor gezogenen 9. Klasse per Zufallsziehung ermittelt.[11]

So wurden n = 6250 Schülerinnen und Schüler für die Testgruppe der Fünfzehnjährigen sowie weitere n = 2989 für die Testgruppe der Neuntklässler ermittelt.[12]

2.2.3. Ziehung der Lehrerstichprobe

Auch die Ziehung der Lehrkräfte erfolgte innerhalb der für PISA 2018 ausgewählten Schulen.

---

[10] Die Ordnungsnummern waren auch in den Listen, welche den Schulen vorlagen, enthalten und dem entsprechenden Namen zugeordnet. Somit erfolgte jegliche Kommunikation zwischen den Schulen und der IEA Hamburg ausschließlich über diese Ordnungsnummern.

[11] Für die Berufsschulen war die Klassenziehung nicht relevant, da an diesen Schulen nur die Testgruppe der fünfzehnjährigen Schülerinnen und Schüler gezogen wurde.

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

Ebenso wie für die Ermittlung der Stichproben auf Schülerebene wurde für die Ziehung der Lehrkräfte von den jeweiligen Schulen eine Liste erstellt, in diesem Fall die Lehrerliste, welche der IEA ebenfalls in pseudonymisierter Form vorlag. Diese enthält wie auch die Schülerliste demografische Merkmale sowie Angaben zu den unterrichteten Fächern. Auch diese Liste wurde in das Stichprobenziehungsprogramm KeyQuest importiert, sodass anschließend die Stichprobenziehung gemäß dem bereits vorgestellten Stichprobendesign für Lehrkräfte durchgeführt werden konnte.

Durch dieses Verfahren ergab sich eine Gesamtzahl von *n* = 2866 Deutschlehrkräften. Davon wurden *n* = 2764 zufällig gezogen und *n* = 102 weitere Deutschlehrkräfte zusätzlich in die Stichprobe aufgenommen, da sie die gezogene 9. Klasse unterrichten. Bei den sonstigen Lehrkräften ergab sich eine Gesamtzahl von *n* = 4800, von denen *n* = 3949 zufällig gezogen wurden sowie *n* = 851 zusätzlich aufgenommen wurden. Damit lag die Gesamtzahl der Lehrerstichprobe für PISA 2018 bei *n* = 7666.

Schließlich wurden die Ergebnisse der Klassen-, Schüler- und Lehrerstichprobenziehungen an die Schulen weitergegeben, sodass diese die Durchführung der Testung vorbereiten konnten.

2.3. PISA Teilnahmequoten (realisierte Stichproben)

Bei den bisher beschriebenen Stichproben handelt es sich um sogenannte Bruttostichproben. Diese Bruttostichproben umfassen alle für die Teilnahme an PISA 2018 ausgewählten Schulen, Schülerinnen und Schüler sowie Lehrkräfte. Hiervon zu unterscheiden sind die Nettostichproben, welche lediglich die Schulen und Personengruppen umfassen, die auch tatsächlich an der Studie teilgenommen haben und von denen auswertbare Daten vorliegen.

Die Brutto- und Nettostichproben in PISA 2018 unterscheiden sich wie folgt: An sechs von 234 insgesamt gezogenen Schulen gab es keine fünfzehnjährigen Schülerinnen und Schüler, sodass keine Testung stattfinden konnte. Bei diesen sechs Schulen handelt es sich um drei Berufsschulen, eine Förderschule, sowie zwei Schulen, die in Grundschulen umgewandelt wurden.[13] Gemäß der internationalen PISA-Standards dürfen solche Schulen nicht durch andere ersetzt werden, da sie für PISA nicht teilnahmeberechtigt sind. [14] Sie haben keinen Einfluss auf die Teilnahmeraten. Zwei weitere Schulen wurden im Verlauf der Studie

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

[14] Gemäß der internationalen PISA-Standards sind für jede gezogene Originalschule zwei Ersatzschulen vorgesehen, die jedoch nur unter bestimmten Bedingungen zum Einsatz kommen können.

ausgeschlossen und ebenfalls nicht ersetzt. Weiterhin verweigerte eine Schule die Teilnahme und konnte durch keine ihrer beiden Ersatzschulen ersetzt werden. Somit ist auch diese Schule in der Nettostichprobe nicht enthalten. Fünf Schulen wurden hingegen jeweils durch ihre erste Ersatzschule ersetzt und eine Schule wurde durch ihre zweite Ersatzschule ersetzt. Werden also von der Bruttostichprobe $n = 234$ die acht nicht teilnahmeberechtigten Schulen abgezogen, ergibt sich eine korrigierte Bruttoschulstichprobe von $n = 226$ Schulen. Hiervon haben 223 Schulen[15] teilgenommen. Dies entspricht einer ungewichteten Teilnahmerate von 95.1 Prozent basierend auf den original gezogenen Schulen und 97.8 Prozent nach Berücksichtigung der Ersatzschulen.

Die gewichtete Teilnahmerate auf Schulebene beträgt 95.7 Prozent ohne und 98.2 Prozent unter Berücksichtigung von Ersatzschulen.

Auch auf der Ebene der Schülerinnen und Schüler kam es zu Ausfällen. Hierfür gibt es verschiedene Gründe. So kann es vorkommen, dass einzelne Jugendliche aufgrund eines sonderpädagogischen Förderbedarfs offiziell von der Studie ausgeschlossen wurden. Da ein solcher Ausschluss bereits vor der Testung erfolgt, werden diese Ausfälle in der Teilnahmequote nicht berücksichtigt. Weitere Ausfälle sind durch Erkrankungen der Schülerinnen und Schüler oder kurzfristig erfolgte Schulwechsel zu erklären.[16] Damit ergibt sich eine korrigierte Bruttoschülerstichprobe von $n = 6056$ für die Testgruppe der fünfzehnjährigen gezogenen Schülerinnen und Schüler. Davon nahmen insgesamt $n = 5451$ teil[17], woraus sich eine ungewichtete Teilnahmerate von 90.0 Prozent ergibt.

Die gewichtete Teilnahmequote auf Schülerebene liegt bei 90.4 Prozent.

Diese hohen Teilnahmequoten lassen auf hoch valide und vollumfänglich repräsentative Ergebnisse schließen.

---

[15] Von diesen 223 Schulen wurden zwei nicht in der Berechnung des Teilnahmestatus berücksichtigt, da die Teilnahme auf Individualebene zu gering war (Teilnahmequote 25% -50%). Diese beiden Schulen sowie die dazugehörigen Schülerinnen und Schüler sind aber trotzdem in den Analysen enthalten.

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

[17] Von diesen 5451 Schülerinnen und Schülern wurden 20 nicht in der Berechnung des Teilnahmestatus berücksichtigt, da sie aus den beiden Schulen mit zu niedriger Teilnahme auf Individualebene (Teilnahmequote 25% -50%) stammen.

Auf die zusätzliche Testgruppe der Neuntklässler wird – wie bereits erwähnt – in diesem Bericht nicht weiter eingegangen.

Da der Lehrerfragebogen nicht in allen Bundesländern verpflichtend war, ist hier eine geringere Teilnahmequote als bei den Schülerinnen und Schülern zu beobachten. Von den insgesamt *n* = 7666 für PISA 2018 ausgewählten Lehrkräften nahmen *n* = 5673 an der Befragung teil. Die ungewichtete Teilnahmequote liegt damit bei 74.0 Prozent, und somit immer noch auf akzeptablem Niveau. Wird dabei zwischen den Quoten der Deutsch- und sonstigen Fachlehrkräfte differenziert, fallen keine wesentlichen Unterschiede auf. Die ungewichtete Teilnahmequote der Deutschlehrkräfte liegt bei 75.1 Prozent, die der sonstigen Lehrkräfte bei 73.4 Prozent.

2.4. Gewichtung als Adjustierung unterschiedlicher Ziehungswahrscheinlichkeiten

Aufgrund des Designs der Stichprobenziehung (mehrstufig) haben nicht alle Schülerinnen und Schüler die gleiche Wahrscheinlichkeit, in die Stichprobe gezogen zu werden. Dazu kommt, dass nicht alle gezogenen Schülerinnen und Schüler auch tatsächlich an der Testung teilnehmen (vgl. 2.3. Realisierte Stichproben). Das Vorliegen gleicher Ziehungswahrscheinlichkeiten für jede Untersuchungseinheit (Schülerinnen und Schüler) und die Teilnahme bei erfolgter Ziehung in die Stichprobe sind aber notwendige Voraussetzungen für die Verallgemeinerbarkeit von Stichprobenergebnissen auf die Zielpopulation (Bortz & Döring, 2006).

Um die ungleichen Ziehungswahrscheinlichkeiten sowie auch unterschiedliche Teilnahmeraten auszugleichen, werden Schulbasis- beziehungsweise Schülerbasisgewichte sowie verschiedene Korrekturfaktoren verwendet. Die Schulbasisgewichte werden umgekehrt proportional zur Ziehungswahrscheinlichkeit errechnet. Haben also beispielsweise Schulen einer Schulform aufgrund des Stichprobendesigns eine geringere Wahrscheinlichkeit, in die Stichprobe gezogen zu werden, ergibt sich entsprechend für diese Schulen ein höheres Schulbasisgewicht. Die Schülerbasisgewichte innerhalb der gezogenen Schulen werden ebenfalls umgekehrt proportional zur Schülerziehungswahrscheinlichkeit errechnet. Hier haben Schülerinnen und Schüler einer großen Schule eine geringere Wahrscheinlichkeit, in die Stichprobe gezogen zu werden, als Schülerinnen und Schüler einer kleinen Schule. Daher erhalten die Jugendlichen, welche eine große Schule besuchen, ein höheres Schülerbasisgewicht.

Weiterhin gehen fünf Korrekturfaktoren in die Gewichtung ein. Zunächst muss der Ausfall von Schulen berücksichtigt werden. Sollte es zu einem Schulausfall kommen, werden andere Schulen, welche derselben expliziten Schicht angehören und somit der ausgefallenen Schule möglichst ähnlich sind, höher gewichtet, um den Ausfall zu kompensieren (1. Korrekturfaktor). Auch auf der Ebene der Schülerinnen und Schüler wird das Basisgewicht um deren Nichtteilnahme korrigiert (2. Korrekturfaktor). Damit wird vermieden, dass es zu einer Über- oder Unterrepräsentation von Schülerinnen und Schülern bestimmter Subpopulationen kommt. Zwei weitere Korrekturfaktoren gleichen Differenzen der Schulbasis- beziehungsweise Schülerbasisgewichte zwischen der Stichprobenziehung und der tatsächlichen Größe der Ziehung zum Zeitpunkt der Erhebung aus. Ein weiterer Korrekturfaktor betrifft Staaten, in denen nur die fünfzehnjährigen Schülerinnen und Schüler befragt werden, welche sich in der Klassenstufe mit der am höchsten zu erwartenden Anzahl an Fünfzehnjährigen befinden. Detaillierte Angaben zur Berechnung der Gewichte können dem Technical Report zu PISA 2015 (OECD, 2017) oder dem Technical Report der OECD zu PISA 2018 entnommen werden (OECD, in Vorbereitung).

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

3. PISA-Daten und Analysegrundlage

Die *nationalen Projektmanagerinnen* (NPMs) bestimmen ein *nationales Datenmanagement Team* (NDM), welches für die termingerechte Bearbeitung aller anfallenden Prozesse des Datenmanagements verantwortlich ist. In Deutschland arbeiten neben dem Datenmanagement des nationalen Projektteams am *Zentrum für internationale Bildungsvergleichsstudien* (ZIB) an der *TUM School of Education der Technischen Universität München* noch Mitarbeiterinnen und Mitarbeiter des Datenmanagements sowie der Kodierabteilung der IEA Hamburg an diesen Aufgaben.

3.1. Datenmanagement

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

Noch bevor die eigentlichen Testungen in den PISA-Schulen durchgeführt werden, wird die internationale Datenmaske um nationale Adaptionen und Ergänzungen der Fragebögen angepasst.

Nationale Adaptionen bezeichnen dabei nationale Änderungen in Antwortformaten, welche eindeutig einer internationalen Kategorie zugeordnet werden können. Ein Beispiel dafür sind die unterschiedlichen Schularten in den teilnehmenden Staaten. Diese Unterschiede werden nach nationaler Erfassung nach der *International Standard Classification of Education* (ISCED-97) entsprechend in internationalen Pendants rekodiert (OECD, 1999). Nationale Ergänzungen können nicht international rekodiert werden. Die Skalenhandbüchern der jeweiligen PISA-Testungen enthalten entsprechende Übersichten (vgl. Mang et al., 2018 und 2019).
Die Adaptionen und Ergänzungen beziehen sich nur auf die nach der eigentlichen Testung anschließenden Fragebögen für Schülerinnen und Schüler beziehungsweise für die Fragebögen der Schulleiterinnen und Schulleiter, der Lehrkräfte und der Eltern. Diesen Schritt bezeichnet man als Harmonisierung der nationalen Datenmaske an ihr internationales Pendant. Die eigentlichen Testfragen werden in ihrer Datenstruktur nicht verändert, so dass hier keine nationalen Anpassungen der Datenmaske notwendig sind. Final liegt demnach noch vor dem eigentlichen Testzeitraum eine an die deutschen Testinstrumente angepasste Datenmaske vor.

Nach der eigentlichen PISA Befragung werden die Test- und Fragebogenantworten der Schülerinnen und Schüler, sowie die beantworteten Fragebögen von Schulleiterinnen und Schulleitern, Lehrkräften und Eltern in einem streng gesicherten Verfahren in das vom internationalen Konsortium zur Verfügung gestellte Datenbanksystem *IEA Data Management Expert* (DME) eingelesen. Die DME-Software ist eine leistungsstarke, eigenständige Anwendung, die auf den meisten Windows-Systemen installiert werden kann und keine Verbindung zum Internet benötigt. Gearbeitet wird auf einer separaten Datenbankdatei mit klar definierten Datensätzen, die den verschiedenen Instrumenten der Studie zugeordnet sind. Der Vorteil der Software liegt auch im Auslesen der Systemdateien, welche neben den Antworten digital erfasst werden, und zum Beispiel die Bearbeitungszeit einer Schülerin oder eines Schülers enthalten.

Eine Vielzahl von Datenbereinigungsschritten und Validitätsprüfungen mit Hilfe sogenannter Konsistenz- und Validitätskontrollen folgen in einem Zeitraum von circa drei Monaten sowohl auf Seite des nationalen Datenmanagement Teams als auch auf Seite des

internationalen Konsortiums. Absprachen erfolgen immer im engen und regelmäßigen Austausch beider Parteien. Das internationale Konsortium stellt im Sommer des Jahres der Berichtslegung die aufbereiteten und mit Schätzwerten der Kompetenzen und weiteren Skalen versehenen Datensätze dem nationalen Projektteam zur internen Nutzung zur Verfügung. Neben diesen Datensätzen gibt es weitere Berichte mit Informationen zur Qualitätssicherung sowie Validierung. Erst mit dem Stichtag der Veröffentlichung der Berichtserstattung werden die internationalen Daten aller beteiligten Staaten über die Webseite der OECD zur freien Nutzung zur Verfügung gestellt. Die nationalen, deutschen Daten werden nach Berichterstattung über das Forschungsdatenzentrum (FDZ) des Instituts für Qualitätsentwicklung im Bildungswesen (IQB) den deutschsprachigen wissenschaftlichen Nutzern freigegeben (vgl. Prenzel et al., 2015; Reiss et al., 2019).

3.2. PISA-Kompetenz

In der PISA-Erhebung 2018 bildet Lesen die Hauptdomäne und Mathematik sowie Naturwissenschaften die Nebendomänen. Deutschland nahm nicht an der Testung der innovativen Domäne Global Competence sowie an der Testung der weiteren optionalen Domäne Financial Literacy teil (OECD, 2019). Unterschiede zwischen Hauptdomäne und Nebendomäne beziehungsweise innovativer Domäne finden sich für PISA 2018 vor allem im Testdesign, welches nachfolgend erläutert wird.

3.2.1. PISA-Testdesign

Nachdem bei PISA 2015 die Umstellung von papierbasierter auf computerbasierte Testung erfolgte (Heine et al., 2016), wurde bei PISA 2018 zum ersten Mal für die Hauptdomäne Lesen das adaptive Testen eingeführt (OECD, 2013). Adaptives Testen impliziert, dass eine Schülerin beziehungsweise ein Schüler je nach Richtigkeit der beantworteten Frage eine schwierigere oder leichtere Folgefrage zur weiteren Beantwortung erhält. Ziel des adaptiven Testens ist das präzisere Erfassen der Fähigkeiten jeder Schülerin beziehungsweise jedes Schülers bei Benutzung einer begrenzten Anzahl an Fragen (van den Linden & Glas, 2010). Für die Nebendomänen Mathematik und Naturwissenschaften wurden die gleichen Items wie in PISA 2015 eingesetzt. Insgesamt gab es in Deutschland 36 Testhefte, sogenannte „forms“, welche aus unterschiedlichen Zusammensetzungen der domänenspezifischen Fragen bestehen und auf die teilnehmenden Schülerinnen und Schüler in folgender Aufteilung zugeordnet werden:

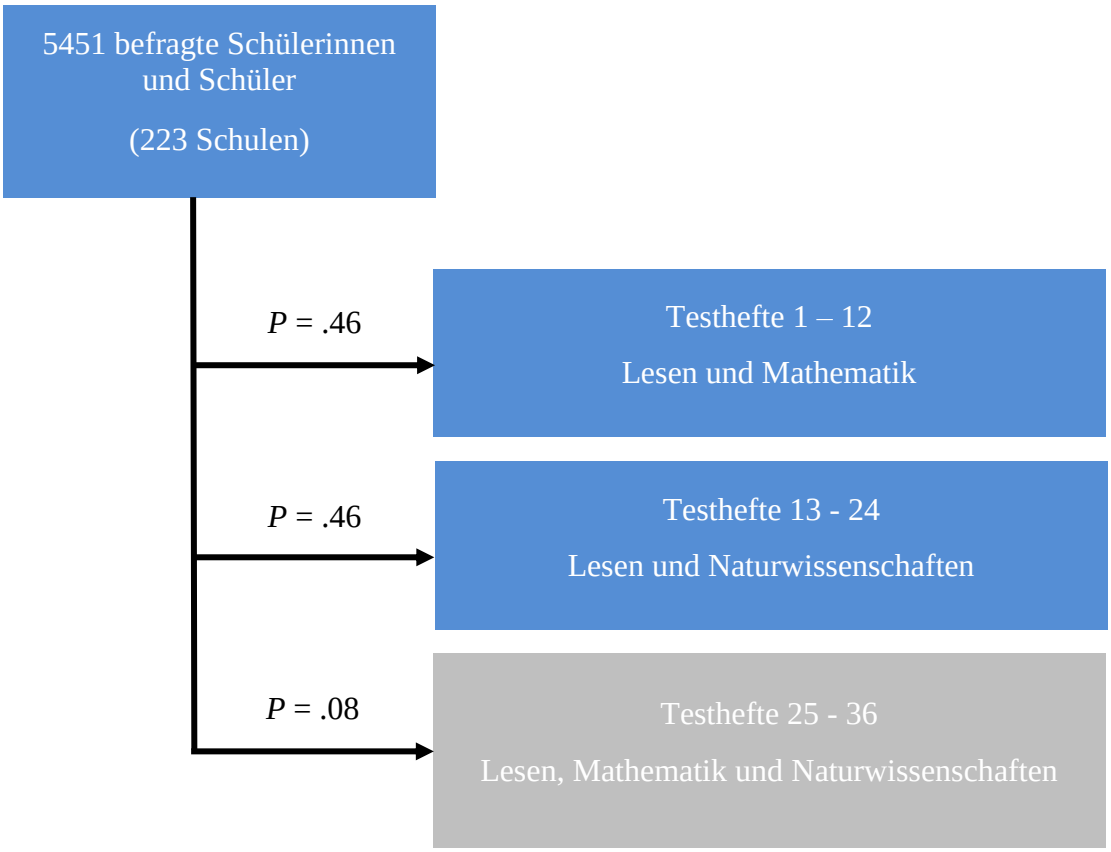

Abbildung 3: PISA 2018 Testheftdesign, adaptiert nach OECD (in Vorbereitung)

Ein Testheft besteht aus vier Clustern, welche aus mehreren Aufgabeneinheiten mit unterschiedlichen Items bestehen.

92 Prozent der Schülerinnen und Schüler erhielten ein Testheft, welches aus vier 30-minütigen Clustern mit zwei Domänen bestand. Insgesamt betrug hier die Testzeit zwei Stunden mit je einer Stunde pro Domäne. Acht Prozent der Jugendlichen erhielten ein Testheft, welches aus vier 30-minütigen Clustern mit drei Domänen bestand. Die gesamte Testzeit betrug ebenfalls zwei Stunden (30-min Cluster je Nebendomäne, zwei 30-min Cluster für die Hauptdomäne Lesen). Zwei je 30-minütige Cluster für die Hauptdomäne Lesen waren in jedem der Testhefte enthalten.

Tabelle 2: Testformen nach Clustern, adaptiert nach OECD (in Vorbereitung)

Testform	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	R		M1	M2
2	R		M2	M3
3	R		M3	M4
4	R		M4	M5
5	R		M5	M6a/b
6	R		M6a/b	M1
7	M1	M3	R	
8	M2	M4	R	
9	M3	M5	R	
10	M4	M6a/b	R	
11	M5	M1	R	
12	M6a/b	M2	R	
13	R		S1	S2
14	R		S2	S3
15	R		S3	S4
16	R		S4	S5
17	R		S5	S6
18	R		S6	S1
19	S1	S3	R	
20	S2	S4	R	
21	S3	S5	R	
22	S4	S6	R	
23	S5	S1	R	
24	S6	S2	R	
25	R		S1	M1
26	R		M2	S2
27	R		S3	M3
28	R		M4	S4
29	R		S5	M5
30	R		M6a/b	S6
31	M1	S1	R	
32	S2	M2	R	
33	M3	S3	R	
34	S4	M4	R	
35	M5	S5	R	
36	S6	M6a/b	R	

Anmerkung: R - Lesen; S - Naturwissenschaften; M - Mathematik

Das in PISA eingesetzte Verfahren des adaptiven Testens wird als *Multistage Adaptives Testen* (MAT) bezeichnet. Das Design besteht aus drei Stufen in welchen die Schülerinnen und Schüler Fragen unterschiedlichen Schwierigkeitsgrades bearbeiten. Die Grundstufe wird als *Core* bezeichnet. Am Anfang der Befragung, wird jede Schülerin und jeder Schüler

zufällig einer Stufe zugeordnet. Je nach Richtigkeit der Antworten, wird der Schüler dann über Stufe 1 und 2 navigiert.

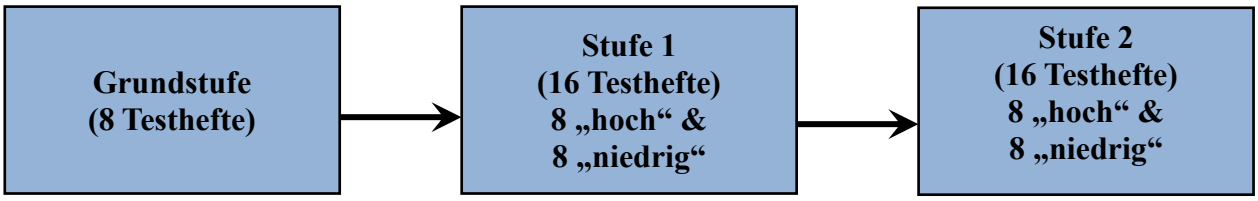

Abbildung 4: PISA 2018 adaptives Testdesign, adaptiert nach OECD (in Vorbereitung)

Insgesamt besteht der gesamte Item-Pool für die Hauptdomäne Lesen aus 245 Items. Jede Schülerin beziehungsweise jeder Schüler beantwortet zwei Aufgabeneinheiten auf der Core Ebene, drei Aufgabeneinheiten auf Ebene 1 und zwei Aufgabeneinheiten auf Ebene 2. Da die entsprechenden Aufgabeneinheiten in ihrer Anzahl der Items variieren, gibt es in Summe pro Jugendlichen zwischen 33 und 40 (durchschnittlich 35-37) zu bearbeitende Items für dieses Cluster. Diese sind pro Stufe und Schwierigkeit in je acht sogenannte „Testlets“ organisiert. Zusätzlich werden je Stufe parallele Testlets eingeführt. Diese sichern die Verbindung beziehungsweise Überschneidung der Schwierigkeitsgrade innerhalb einer Stufe.

Die Entscheidung, welche Folge-Aufgabeneinheiten die Schülerinnen und Schüler erhalten, wird nach drei Kriterien entschieden:

1. Die zugewiesenen Stufen für Core beziehungsweise Stufe 1.
2. Die Leistung des Jugendlichen in der jeweiligen Frage (die Anzahl der richtig beantworteten Items in dieser Stufe).
3. Eine Wahrscheinlichkeitsmatrix, mit welcher sichergestellt wird, dass alle Testlets auch zum Einsatz kommen und keine Ausreißer aufkommen können.

Die Leistung eines Jugendlichen auf der jeweiligen Stufe wird dabei zum einen über richtig beantwortete Items der vorherigen Stufe über PISA-Schwellenwerte in „niedrig“, „mittel“ und „hoch“ eingeteilt. Die PISA-Schwellenwerte definieren sich aus der Anzahl der richtig beantworteten Items, um einen Kompetenzwert von 425 (Übergang von der niedrigen zur mittleren Kategorie) sowie um einen Kompetenzwert von 530 (Übergang von der mittleren zur hohen Kategorie) zu erhalten. Zum anderen basiert die Entscheidung auf einer durchschnittlichen Wahrscheinlichkeit für ein bestimmtes Testlet alle Items mit diesen vordefinierten Kompetenzen richtig zu beantworten. Beide Schritte optimieren das adaptive Testdesign, so dass die limitierte Information jeder Stufe (nur automatisch kodierte Items können verwendet werden) bestmöglich ausgewertet werden kann.

Um fehlende Werte auf Stufe 2 aufgrund von Müdigkeitseffekten oder ähnlichen Gründen zu vermeiden, wird bei 25 Prozent der befragten Schülerinnen und Schülern die Stufenzuordnung vertauscht, das heißt Stufe 1 folgt nach Stufe 2 und Core. Die restlichen 75 Prozent der Jugendlichen behalten die reguläre Reihenfolge mit Stufe 2 nach Stufe 1 und Core.

3.2.2. Kodierung offener Testantworten

Im Gegensatz zu geschlossenen Antwortformaten wie Mehrfach-Wahlaufgaben (Multiple-Choice-Aufgaben), die ohne großen Aufwand verwertbare Daten liefern, erfordern sogenannte offene Antwortformate, bei denen die Schülerinnen und Schüler angehalten werden, etwa mit einem Antwortsatz sprachproduktiv zu reagieren, einen zusätzlichen Schritt zur quantifizierbaren Auswertung. Geschulte Auswerterinnen und Auswerter überführen diese offenen Antworten in ein geschlossenes System, in der Regel mit Kategorien wie „richtige Antwort“, „teilweise richtige Antwort“ und „falsche Antwort“. Da es in der Natur authentischer Sprachproduktion liegt, dass die Beurteilung immer eine subjektive Restkomponente enthält, erfolgt die Bewertung anhand fester Kriterien, die in einer verbindlichen Kodieranweisung festgelegt sind. Dieser Kriterienkatalog verfolgt, wie das weitere Prozedere des Kodierungsprozesses, das vorrangige Ziel, den Ermessensspielraum der Auswerterinnen und Auswerter zu minimieren und so eine möglichst objektive und einheitliche Bewertung zu ermöglichen.

Bei PISA 2018 gibt es insgesamt 135 solcher Aufgaben mit offenem Antwortformat, verteilt auf die drei Domänen Lesen (85 Aufgaben), Mathematik (18 Aufgaben) und Naturwissenschaften (32 Aufgaben). Für die deutsche Stichprobe kamen für die Domäne Lesen acht und für Mathematik und Naturwissenschaften jeweils sechs Auswerterinnen und Auswerter zum Einsatz.

Im Vorfeld der Kodierung wurden die Auswerterinnen und Auswerter in drei ganztägigen Trainingssitzungen, verteilt über mehrere Wochen, geschult. Nach jeder der Sitzungen bearbeiteten sie beispielhafte, authentische Schülerantworten, um den Schulungsfortschritt zu dokumentieren und analytisch ermitteln zu können, welche Aufgaben noch Schwierigkeiten bereiten und im weiteren Schulungsverlauf besondere Aufmerksamkeit erfordern. Für die Domäne Lesen vergab jede Auswerterin und jeder Auswerter während der Schulungsphase 9600 Übungscodes, in der Mathematik 1570 und in der Naturwissenschaft 3460. Die Kodierung erfolgte dabei über den *Coding Expert Client* der IEA, der über ein Trainingsmodul verfügt, das während der Schulungen zum Einsatz kam.

Die anschließende Kodierung der offenen Schülerantworten erfolgte über das *Open-Ended Coding System* (OECS; OECD, 2017); diese webbasierte Plattform kam erstmals bei PISA 2015 zum Einsatz und automatisiert die Verteilung der Schülerantworten an die Auswerterinnen und Auswerter.

Zur Ermittlung der Zuverlässigkeit der Kodierung innerhalb eines teilnehmenden Staates ist es erforderlich, dass ein Teil der Antworten mehrfach bewertet wird (vgl. ETS, 2018). Hierfür werden die Antworten jeweils zwei Personen zugeteilt, die unabhängig voneinander die Kodierung vornehmen. Anhand der Abweichung oder Nichtabweichung dieser vergebenen Codes kann ein Übereinstimmungswert ermittelt werden, der einen Rückschluss auf die Reliabilität der Kodierung zulässt. Als Zielsetzung geben die *PISA 2018 Main Survey Coding Procedures (CBA Countries)* eine Übereinstimmung von 92 Prozent über alle Items hinweg vor. Für einzelne Items gilt eine Übereinstimmung von 85 Prozent als zufriedenstellend.

Ein ähnliches Verfahren wird eingesetzt, um die Reliabilität zwischen den teilnehmenden Staaten zu ermitteln. Hierfür wird in jedem Staat ein identisches Kontingent von englischsprachigen Antworten kodiert, die dann zwischen den Staaten verglichen werden können.

Erstmals bei PISA 2018 ist ein Teil der Antworten automatisiert vom OECS ausgewertet worden. Dazu wurden Antworten aus vorhergehenden PISA-Zyklen ausgewertet und besonders häufig, eindeutige bewertete Codes auf entsprechende Antworten aus PISA 2018 übertragen. Insbesondere betrifft das zum Beispiel leere, also vom Schüler nicht bearbeitete Antworten, kurze Ein-Wort-Antworten sowie – insbesondere in der Domäne Mathematik – Antworten in Form einer einzelnen Zahl.

Die Zuordnung von Schülerantworten an Auswerterinnen und Auswerter, einschließlich der Doppelkodierung zur Ermittlung von Reliabilitäten, wird automatisch durch das OECS implementiert. Ebenso ermittelt das OECS nach Abschluss der Kodierung die Reliabilitäten.

Im Mittel liegt die prozentuale Übereinstimmung bei der Bewertung der Aufgaben für die Teilstichprobe in Deutschland bei 92 Prozent. Im Bereich Lesen wurde dabei eine Reliabilität von 93 Prozent erreicht, im Bereich Naturwissenschaft waren es 89 Prozent und im Bereich Mathematik 95 Prozent. Bei der Domäne Lesen lagen drei der 85 Aufgaben unter der Schwelle von 85 Prozent, der niedrigste Wert war dabei 83 Prozent. Bei der Naturwissenschaft erreichten sieben der 32 Aufgaben nicht die 85 Prozent-Marke. Die niedrigste Übereinstimmung eines einzelnen Items lag bei 69 Prozent (nächstniedrige 77

Prozent). Bei der Mathematik lag die niedrigste Übereinstimmung eines Einzelitems bei 88 Prozent. Der Bereich der Mathematik erzielte damit insgesamt die höchste Übereinstimmung. Das vergleichsweise niedrigere Ergebnis der Domäne Naturwissenschaft erklärt sich in großen Teilen dadurch, dass bei PISA 2015 für diesen Bereich viele neue Items entwickelt wurden, deren Kodieranweisungen mit Hilfe der zukünftigen Erhebungen noch optimiert werden sollten.

3.2.3. Kodierung von Berufsangaben nach ISCO-08

Neben den Antworten aus den Testheften werden bei PISA 2018 auch Angaben zum sozioökonomischen Hintergrund der Schülerinnen und Schüler erhoben. Besondere Beachtung verdienen dabei die Angaben zum Beruf der Eltern und zum Wunschberuf der Schülerin oder des Schülers, die – vergleichbar mit dem Prozedere zur Kodierung der offenen Testantworten – in ein geschlossenes System überführt werden müssen, bevor sie ausgewertet werden können. Als Klassifizierungsschema wird dabei die *International Standard Classification of Occupations* in der aktuellen Revision von 2008 (ISCO-08) benutzt, die vom International Labour Office entworfen wurde und international für solche Fälle zum Einsatz kommt (ILO, 2012). Die ISCO-08 unterteilt Berufe nach Tätigkeitsfeldern in 10 Hauptgruppen wie Führungskräfte, Technikerinnen und Techniker sowie Handwerkerinnen und Handwerker. In weiteren Hierarchiestufen finden sich 43 Berufsgruppen, 130 Berufsuntergruppen sowie auf der detailliertesten Ebene 436 Berufsgattungen. Aus den vier Hierarchieebenen ergibt sich so ein ein- bis vierstelliger Code, der die Berufsangabe widerspiegelt. Ziel der Kodierung ist es, anhand der vorhandenen Angaben eine möglichst genaue Klassifizierung vorzunehmen, das heißt nach Möglichkeit bis zur untersten Ebene vorzudringen.

Im Gegensatz zu den Antworten der Testhefte erfolgt die Kodierung der Berufsangaben nicht über das OECS, sondern über den *Coding Expert Client* der IEA. Durchgeführt wurde die Bewertung von einem Team aus sechs Auswerterinnen und Auswertern, die bereits in anderen Studien umfangreiche Erfahrungen mit der Berufskodierung gesammelt haben. Bei PISA 2018 wurden insgesamt 22035 Berufsangaben kodiert.

3.2.4. Statistische Berechnungsverfahren der PISA-Kompetenz

Ein zentraler Aspekt der PISA-Studie ist es, die Kompetenzen der Schülerinnen und Schüler innerhalb jedes teilnehmenden Staates zu erfassen und im nationalen und internationalen Kontext zu analysieren und zu vergleichen. Diese Kompetenzen werden mit Hilfe statistischer

Schätzverfahren aus den Antworten der Jugendlichen der PISA-Testung errechnet beziehungsweise geschätzt.

Das Testdesign von PISA beinhaltet, dass nicht jede Schülerin beziehungsweise jeder Schüler alle Testaufgaben beantwortet, sondern nur einen Teil davon. Daher kann man Aussagen über die gemeinsamen Kompetenzen dieser Schülerinnen und Schüler nicht auf prozentualer Basis der richtig oder falsch beantworteten Fragen vornehmen. Des Weiteren können Punktschätzungen, welche für die Kompetenzschätzung einer einzelnen Schülerin beziehungsweise eines einzelnen Schülers durchaus geeignet erscheinen, starke Verzerrungen der Schätzungen für die Gesamtheit der PISA Jugendlichen in einem Staat mit sich ziehen (Wingersky, Kaplan & Beaton, 1987). Die Grundlagen der statistischen Berechnungsverfahren in PISA können unter anderem in Adams, Wilson, Glas und Verhelst (1995), Mislevy und Sheehan (1987), Yamamoto und Mazzeo (1992) sowie Wu, Adams und Wilson (1997) im Detail nachgelesen werden. Aktuellere Übersichten der verschiedenen Aspekte der Methodik finden sich auch in Glas und Jehangir (2014), Mazzeo und von Davier (2014), von Davier und Sinharay (2014), Weeks, Yamamoto und von Davier (2014) und von Davier, Sinharay und Oranje (2006).

Die verwendeten Schätzmethoden in PISA basieren auf Modellen, die ursprünglich im Rahmen der *Item Response Theorie* (IRT; Rasch, 1960) entwickelt wurden und sehr flexibel für die Auswertung von Studien wie PISA anwendbar sind (siehe dazu Skrondal & Rabe-Hesketh, 2004; von Davier & Yamamoto, 2004, 2007; Adams, Wu & Carstensen, 2007). Die Grundlage der IRT besteht aus der Schätzung von Item und Personen Parameter. Die Item Parameter sind zum Beispiel Item-Schwierigkeit, Item-Diskrimination, Ratewarscheinlichkeit, die Personenparameter werden häufig auch als Fähigkeitsparameter bezeichnet. Mit IRT Modellen wird das Antworten auf ein Item als Wahrscheinlichkeitsfunktion der Personen- und Itemmerkmale modelliert.

Ein Postulat dieser Theorie ist die Unabhängigkeit der Personen und Item Parameter. Im Detail bedeutet dies, dass die Item-Schwierigkeit unabhängig von der Gruppe befragter Schülerinnen und Schüler geschätzt wird. Umgekehrt kann die Personenfähigkeit unabhängig von den verwendeten Testaufgaben geschätzt werden. Das in PISA verwendete Testdesign wird daher sehr passend durch diese Theorie abgedeckt.

Das Populationsmodell (von Davier & Sinharay, 2013) zur Schätzung der Kompetenzen in PISA setzt sich aus zwei Schritten zusammen:

1. Item Parameter werden mit Hilfe von IRT Modellen geschätzt.
2. Personenfähigkeiten, die sogenannten Kompetenzen der Schülerinnen und Schüler, werden mit einem latenten Regressionsmodell geschätzt.

Für die Schätzung der Item Parameter wird unterschieden, ob ein Item eine „Ja/Nein" Antwort (dichotom) oder eine mehrstufige Skala (zum Beispiel „Stimmt genau/Stimmt etwas/Stimmt eher nicht/Stimmt überhaupt nicht) besitzt. Für dichotome Items werden die dazugehörigen Item-Schwierigkeiten und Item-Diskriminationen mit Hilfe des *two-parameter logistic Modells* (2PLM, Birnbaum, 1968) geschätzt. Besitzt das Item eine mehrstufige Skala wird das *generalised partial credit Modell* (Muraki, 1992) zur Schätzung der Diskrimination und der Schwellenparameter eingesetzt. In der PISA-Erhebung des Jahres 2015 lösten diese Modelle das klassische *Rasch Modell* (Rasch, 1960) sowie das *partial credit Modell* (Masters, 1982) ab (entsprechend dem dichotomen oder mehrstufigen Antwortformat). Eine detaillierte Auflistung dieser Umstellung findest sich zum Beispiel in Heine et al. (2016) sowie dem Technical Report PISA 2015 (OECD, 2017).

Diese Item-Parameter werden dann mit Hilfe der Testantworten der Schülerinnen und Schüler aller teilnehmenden Staaten und über alle PISA-Erhebungen rückwirkend bis zum Jahre 2009 (hier bildete Lesen das letzte Mal die Hauptdomäne) geschätzt (siehe auch unter 4. Vergleichbarkeit der PISA-Befragungen). Für Items, welche in unterschiedlichen Staaten unterschiedlich schwierig sind, werden die Schwierigkeiten nur für diesen Staat geschätzt und als sogenannte *unique item parameter* bezeichnet und weiter verrechnet (Glas & Jehangir, 2014; Glas & Verhelst, 1995; Yamamoto, 1997; Oliveri & von Davier, 2014, 2011). Detaillierte Modellvoraussetzungen, Grundlagen und Grafiken finden sich ebenfalls im Technical Report PISA 2015 (OECD, 2017) sowie im zu erwartenden Technical Report PISA 2018 (OECD, in Vorbereitung).

Die in Schritt 1 geschätzten Item Parameter werden dann zusammen mit vielen gewonnenen Informationen aus dem Fragebogen für Schülerinnen und Schüler (so zum Beispiel der Zuwanderungshintergrund sowie die sozioökonomische Herkunft eines Jugendlichen) als unabhängige Variablen in Schritt 2 in ein sogenanntes latentes Regressionsmodell aufgenommen. Auch hier steht der Begriff „latent" für eine nicht direkt messbare Größe und bildet kennzeichnende Parameter der Verteilung der Kompetenz einer Schülerin beziehungsweise eines Schülers ab. Anders ausgedrückt werden mit Hilfe dieser Informationen die gesamte Verteilung der möglichen Kompetenzausprägungen eines Jugendlichen abgebildet. Aus dieser Verteilung werden dann 10 *Plausible Values* (PVs;

Mislevy & Sheehan, 1987; von Davier, Gonzalez & Mislevy, 2009) zu besseren Handhabung gezogen und für die weitere Verarbeitung in den öffentlichen Datensätzen zur Verfügung gestellt. Analyseanleitungen und Programmempfehlungen für die Arbeit mit PVs sowie anschauliche Darstellungen finden sich ebenfalls im Technical Report PISA 2015 (OECD, 2017) sowie im Technical Report PISA 2018 (OECD, in Vorbereitung).

3.2.5. Statistische Berechnungsverfahren weiterer PISA Themengebiete

Neben der Schätzung und statistischen Berechnung der PISA-Kompetenzen werden in jeder Erhebung noch weitere sogenannte *Indizes* und *Skalen* als Zusammenschluss mehrerer thematisch zusammengehöriger Fragen des Schüler-, aber auch des Schulleiter-, Eltern- und Lehrerfragebogens errechnet. Diese Skalen bilden nicht direkt beobachtbare - latente - Kenngrößen ab und umfassen Themen wie zum Beispiel den Zuwanderungshintergrund, Motivation und den sozioökonomischen Hintergrund.

Indizes werden mit Hilfe von arithmetischen Transformationen oder Rekodierungen der zugehörigen Fragen gebildet. Skalen basieren, wie bei der Schätzung der PISA-Kompetenzen auf der *Item Response Theorie* (IRT; Rasch, 1960). Dabei werden, analog zu den PISA-Kompetenzen, die Modelle des *two-parameter logistic Modells* (2PLM, Birnbaum, 1968) für dichotome Antwortmuster sowie des *generalised partial credit Modells* (Muraki, 1992) für multivariate Antwortformate verwendet. Alle Antworten der Jugendlichen mit mehr als drei validen Antworten werden in die Schätzung der Skalen eingebunden und mit Gewichten versehen, so dass jeder Staat für die Berechnung zu gleichen Teilen berücksichtigt wird (OECD, 2017).

Es wird ferner unterschieden, ob die Skala eine sogenannte Trendskala ist oder nur in der aktuellen Erhebung erfasst wurde. Wurde die Skala bereits im Jahr 2009 erfragt, als Lesen das letzte Mal die Hauptdomäne bildete, so werden die Parameter der jeweiligen Modelle gemeinsam mit den Antworten der Schülerinnen und Schüler aller teilnehmenden Staaten aus den Jahren 2009 und 2018 geschätzt. Die daraus resultierenden Werte werden nun mittels einer linearen Transformation an die Metrik aus dem Jahre 2009 angepasst. Wurde die Skala nur im Jahre 2018 erhoben, so werden diese Parameter „nur“ mit den Antworten aller Schülerinnen und Schüler aller teilnehmenden Staaten aus PISA 2018 errechnet (OECD, 2017).

Anders als bei der Schätzung der PISA-Kompetenzen werden für die Schätzung des latenten Wertes einer Skala (zum Beispiel der Motivation) keine Regressionen mit

Hintergrundinformationen verwendet. Lediglich ein Schätzwert, der sogenannte *Weighted Likelihood Estimate* (WLE) bestimmt die Ausprägung der latenten Größe. Dieser wird in den meisten Fällen im Anschluss noch an die internationale Metrik der OECD Teilnehmerstaaten mit dem Mittelwert $M = 0$ und einer Standardabweichung $SD = 1$ normiert. Eine detaillierte Beschreibung zu diesem Schätzverfahren findet sich im Technical Report PISA 2015 (OECD, 2017). Eine detaillierte Auflistung aller Indizes und Skalen findet sich des Weiteren in den Skalenhandbüchern der PISA-Erhebungen (vgl. Mang et al, 2018 und 2019).

3.2.6. Statistische Verwertung von Antwortzeiten

Ein großer Vorteil der computerbasierten Befragung ist die Erhebung von Antwort- beziehungsweise Reaktionszeiten. Diese Zeiten werden für jedes Item des Schülerfragebogens in Form von Millisekunden erhoben und den wissenschaftlich interessierten Nutzerinnen und Nutzern auch in den öffentlich verfügbaren Datensätzen zur Verfügung gestellt.

3.2.7. Reliabilität der PISA-Daten

Durch die zweistufige Stichprobenziehung ergibt sich bei allen Analysen eine (statistische) Abhängigkeit der Schülerinnen und Schüler innerhalb der Schulen. Das bedeutet, dass die Merkmale der Jugendlichen (zum Beispiel deren Kompetenzen) innerhalb einer Schule wahrscheinlich ähnlicher ausfallen als zwischen unterschiedlichen Schulen. Des Weiteren unterscheiden sich diese Abhängigkeiten auch zwischen den einzelnen Teilnehmerstaaten (zum Beispiel aufgrund unterschiedlicher Bildungssysteme), was wiederum unterschiedliche Auswirkungen auf die Schätzgenauigkeit haben kann (vgl. OECD, 2017). Der Effekt des Stichprobendesigns auf den Fehler der Populationsschätzungen (dem sogenannten Standardfehler) wird bei PISA und anderen Schulleistungsstudien als *Designeffekt* bezeichnet (Adams, 2005). Dieser Designeffekt wird unter anderem zur Verbesserung der Effizienz der Schätzwerte bei komplexen Stichprobenziehungen eingesetzt, wie sie bei PISA angewendet werden (Cochran, 1977; Rust & Rao, 1996).

Da die Größe des Designeffekts bei PISA im Vorfeld nicht eindeutig zu quantifizieren ist, und darüber hinaus auch zwischen den einzelnen Teilnehmerstaaten unterschiedlich ausfällt, wird der Standardfehler in PISA seit der ersten Erhebung im Jahr 2000 mit sogenannten Replikationsmethoden auf Basis der erhobenen Daten berechnet. Bei PISA kommt die *Balanced Repeated Replication* (BRR, zum Beispiel Wolter, 1985) zur Anwendung, mit einer Erweiterung nach Fay (1989; vgl. dazu auch Judkins, 1990). Die statistische Herleitung der Schätzung des Standardfehlers sowie eine detailliertere Darstellung in Bezug auf die PISA-

Studie ist dem Technical Report zu PISA 2015 (OECD, 2017) zu entnehmen und wird für PISA 2018 im Technical Report der OECD erwartet (OECD, in Vorbereitung).

Durch den Einsatz von Replikationsmethoden werden die Varianzen zwischen und innerhalb der Schulen berücksichtigt. Damit wird einer möglichen Unterschätzung der Standardfehler vorgebeugt (vgl. Heine et al., 2016; OECD, 2017). Bei der Datenauswertung wird die Berechnung der Varianz durch die Verwendung sogenannter Replikationsgewichte im Datensatz praktisch realisiert (OECD, 2017).

3.3.Darstellungsformen der PISA-Kompetenzen

Zur besseren Interpretation der PISA-Kompetenzen der Schülerinnen und Schüler werden diese Werte standardisiert, das heißt so transformiert, dass eine Vergleichbarkeit zwischen unterschiedlichen Staaten oder Gruppen von Schülerinnen und Schülern möglich ist. Dabei definiert sich diese Transformation mit den Kenngrößen $M = 500$ für den Mittelwert und $SD = 100$ für die Standardabweichung bei der erstmaligen Erfassung der jeweiligen Domäne als Schwerpunkt über alle OECD-Staaten hinweg.

Des Weiteren wurden bereits bei PISA 2000 sogenannte Kompetenzstufen je Domäne definiert. Sie bilden ein anschauliches Darstellungsmittel und werden aus den Charakteristiken eines Items und inhaltlich begründeten Mustern gebildet (OECD, 2002).

Je Domäne definieren sich die Stufen mit gleicher Breite auf Basis einer sogenannten „hinreichenden“ Lösungswahrscheinlichkeit. Diese wurde bereits in der PISA-Erhebung des Jahres 2000 auf den numerischen Wert von 62 Prozent festgelegt und bestimmt die Wahrscheinlichkeit ein typisches Item in einer bestimmten Stufe richtig zu lösen. Zusätzlich gilt, dass eine Kompetenzstufe am unteren Ende immer dann erreicht wird, wenn Aufgaben auf dieser Stufe mit einer Wahrscheinlichkeit von mindestens 50 Prozent gelöst werden. Am oberen Ende einer Kompetenzstufe können die Fünfzehnjährigen die Aufgaben mit einer Wahrscheinlichkeit von etwa 70 Prozent richtig lösen (OECD, 2017).

Für den Kompetenzbereich Lesen wurden in PISA 2000 (hier war diese Domäne das erste Mal Hauptdomäne; OECD, 2002) fünf Kompetenzstufen gebildet, welche in PISA 2009 im Bereich der ersten Kompetenzstufe erweitert wurden (OECD, 2012). Des Weiteren wurde in dieser Erhebungsrunde eine weitere Kompetenzstufe am unteren Ende der Skala hinzugefügt (OECD, in Vorbereitung). Für die Bereiche Naturwissenschaften und Mathematik wurden jeweils sechs Kompetenzstufen in den PISA-Erhebungsrunden 2006 und 2003 definiert (OECD, 2009; OECD, 2005), wobei diese für die Domäne Naturwissenschaften in PISA 2015

ebenfalls im unteren Bereich (erste Kompetenzstufe) nochmals differenziert wurden (OECD, 2017). Für die Hauptdomäne Lesen werden noch zusätzliche Kompetenzstufen für die weiteren Teildomänen *Lokalisieren von Informationen*, *Textverstehen* sowie *Bewerten und Reflektieren* definiert.

Eine genaue Angabe der Kompetenzstufen und ihre numerischen Grenzen finden sich im Technical Report für PISA 2018 (OECD, in Vorbereitung).

4. Vergleichbarkeit der PISA-Befragungen

Obwohl die PISA-Erhebungen über alle bisherigen Zyklen immer als querschnittliche Befragung angelegt sind (es beantworten in jeder Erhebung andere Schülerinnen und Schüler die Testfragen), besteht ein großes Interesse daran, Vergleiche für die Kompetenzen zwischen den Erhebungen herzustellen. Neben der unterschiedlichen Stichprobe müssen noch weitere Faktoren berücksichtigt werden (Carstensen, Prenzel & Baumert, 2009). Diese sind mögliche Abweichungen in der Schätzung der Kompetenzen der Jugendlichen je Erhebungsrunde, Unterschiede basierend auf dem Wechsel von papierbasierter zu computerbasierter Erhebung und dem Wechsel zum adaptiven Testen.

Diese Unsicherheiten werden bei PISA mit Hilfe eines *Link Errors* (Rousseeuw & Croux, 1993) ausgeglichen, welcher mit Hilfe von Testaufgaben errechnet wird, die in jeder Erhebung eingesetzt werden (sogenannte *Link Items*). Ab PISA 2015 wurde das Verfahren optimiert, indem die Item Parameter Schätzung nicht mehr nur mit Daten der aktuellen Erhebung, sondern über alle Erhebungen bis zur letzten gleichen Hauptdomäne durchgeführt wurde. Nähere Informationen zu diesem *Linking Design* finden sich zum Beispiel im Technical Report zu PISA 2015 (OECD, 2017).

5. Zusammenfassung und Ausblick

Die PISA-Studie als manifeste Befragung seit nunmehr über 18 Jahren mit sieben Erhebungen passt mit der Einführung des adaptiven Testens für die Hauptdomäne Lesen die Studie an den aktuellen methodischen Status Quo der Wissenschaft an. Aufgrund dieses Schrittes wird bei künftigen Erhebungen nicht mehr in der Anzahl der Items für Haupt- und Nebendomäne unterschieden, sodass Vergleiche über die Erhebungen valider sein dürften. Allerdings sollte die Unsicherheit der Schätzungen mit dem Wechsel auf dieses Design nicht unberücksichtigt bleiben und sowohl in Analysen als auch bei der Interpretation der Ergebnisse einbezogen werden.

Für PISA 2021 wird als nächste Hauptdomäne Mathematik in das Design des adaptiven Testens aufgenommen. Die dann als Nebendomäne deklarierte Domäne Lesen wird weiterhin adaptiv getestet, die Domäne Naturwissenschaften wird im Jahre 2024 in das adaptive Design aufgenommen. Darüber hinaus wird es sogenannte innovative Domänen geben, in denen weitere Herausforderungen an fünfzehnjährigen Schülerinnen und Schüler untersucht werden.

6. Literatur

Adams, R. J., Wilson, M. R. & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioural Statistics, 22,* 46-75.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2–3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, R. J., Wu, M. L. & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In M. Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (S. 271-280). New York, NY: Springer.

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. 395–479). Reading, MA: Addison-Wesley.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Brown, R. S. (2010). Sampling. In P. P. McGaw (Hrsg.), *International encyclopedia of education* (3. Aufl., S. 142–146). Oxford: Elsevier.

Carstensen, C. H., Prenzel, M. & Baumert, J. (2009). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? In M. Prenzel & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006* (S. 11-34). Wiesbaden: VS Verlag für Sozialwissenschaften.

Cochran, W. G. (1977). *Sampling techniques* (3. Aufl.). New York: John Wiley.

ETS. (2018). *PISA 2018 Main Survey Coding Procedures (CBA Countries)* (interner Bericht). Princeton, N.J.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. In *Proceedings of the American Statistical Association* (S. 212–217). Washington, D.C.: Alexandria, VA: American Statistical Association.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models: Foundations, Recent Developments, and Applications* (S. 69-95). New York, NY: Springer.

Glas, C. A. W. & Jehangir, K. (2014). Modelling country-specific differential item functioning. In L. Rutkowski, M. v. Davier & D. Rutkowski (Hrsg.), *Handbook of International Large-Scale Assessment* (S. 97-115). Boca Raton, FL: CRC Press.

Häder, M. (2015). *Empirische Sozialforschung – Eine Einführung* (3. Aufl.). Wiesbaden: Springer VS.

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

ILO (2012). *International Standard Classification of Occupations. Structure, group definitions and correspondence tables.* Genf: International Labour Office. Verfügbar unter *https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf*

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

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

Levy, P. S. & Lemeshaw, S. (2008). *Sampling of populations – methods and applications* (4. Aufl.). Hoboken, NJ: Wiley & Sons.

Mang, J., Ustjanzew, N., Schiepe-Tiska, A., Prenzel, M., Sälzer, C., Müller, K. & González Rodríguez, E. (2018). *PISA 2012 Skalenhandbuch. Dokumentation der Erhebungsinstrumente.* Münster: Waxmann.

Mang, J., Ustjanzew, N., Leßke, I., Schiepe-Tiska, A. & Reiss, K. (2019). *PISA 2015 Skalenhandbuch. Dokumentation der Erhebungsinstrumente.* Münster: Waxmann.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272

Mazzeo, J. & v. Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. v. Davier & D. Rutkowski (Hrsg.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (S. 229-258). Boca Raton, FL: CRC Press.

Mislevy, R. J. & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Hrsg.), *Implementing the new design: the NAEP 1983-84 technical report* (S. 293–360). Princeton, N.J: National Assessment of Educational Progress, ETS.

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 (1999). *Classifying educational programmes. Manual for ISCED- 97 implementation in OECD countries.* Paris: OECD.

OECD (2013). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD.

OECD (2017). *PISA 2015 technical report*. Paris: OECD.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD.

OECD (in Vorbereitung). *PISA 2018 technical report*. Paris: OECD.

Oliveri, M. E. & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modelling, 53*(3), 315-333. Abgerufen unter http://www.psychologie-aktuell.com/index.php?id=200

Prenzel, M., Sälzer, C., Klieme, E., Köller, O., Mang, J., Heine, J.-H., Schiepe-Tiska, A. & Müller, K. (2015): *Programme for International Student Assessment 2012 (PISA 2012).* Version: 3. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2012_v3

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

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Mang, J., Heine, J.-H., Weis, M., Klieme, E. & Köller O. (2019), *Programme for International Student Assessment 2015 (PISA 2015).* Version 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz. http://doi.org/10.5159/IQB_PISA_2015_v1

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*(424), 1273. https://doi.org/10.2307/2291267

Rust, K. & Rao, J. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research, 5*(3), 283–310. https://doi.org/10.1177/096228029600500305

Skinner, C. J. (2014). *Probability proportional to size (PPS) sampling*. Wiley StatsRef: Statistics Reference Online, 1-5.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models.* Boca Raton, FL: Chapman and Hall/CRC.

Statistisches Bundesamt & DESTATIS (2017). *Bildung und Kultur: Allgemeinbildende Schulen. Schuljahr 2017/2018 (Fachserie 11 Reihe 1).* Wiesbaden: Statistisches Bundesamt. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100187004.pdf?__blob=publicationFile.

Thompson, S. K. (2012). *Sampling* (3. Aufl.). Hoboken, NJ: Wiley & Sons.

van der Linden, W. J., Glas, C. A. W. (2010): Elements of Adaptive Testing. New York, NY: Springer Science+Business Media, LLC (Statistics for Social and Behavioral Sciences).

von Davier, M. & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389-406.

von Davier, M., Sinharay, S. & Oranje, A. (2006). The Statistical Procedures Used in National Assessment of Educational Progress: Recent Developments and Future Directions. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics 26: Psychometrics* (S. 1039–1055). Amsterdam, NL: Elsevier.

von Davier, M. & Yamamoto, K. (2007). Chapter 6: Mixture distribution Rasch models and Hybrid Rasch models. In M. von Davier & C.H. Carstensen, *Multivariate and Mixture Distribution Rasch Models.* New York, NY: Springer.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009). What are plausible values and why are they useful? In *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, 9-36. Abgerufen unter http://www.ierinstitute.org/ieri-home.html

von Davier, M. & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook Of International Large-Scale Assessment: Background, Technical Issues, And Methods Of Data Analysis* (S. 155-174). Boca Raton, FL: CRC Press.

Weeks, J., Yamamoto, K. & von Davier, M. (2014). Design considerations for the Program for International Student Assessment. In L. Rutkowski, M. von Davier and D. Rutkowski (Hrsg.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (S. 259-275). Boca Raton, FL: CRC Press.

Wingersky, M., Kaplan, B. & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton (Hrsg.*), Implementing the new design: The NAEP 1983-84 technical report* (S. 285-292). Princeton, NJ: Educational Testing Service.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer.

Wu, M. L., Adams, R. J. & Wilson, M. R. (1997). *ConQuest: Multi-Aspect Test Software* [computer program]. Camberwell, Australia: Australian Council for Educational Research.

Yamamoto, K. & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155-174.

Yamamoto, K. (1997). Scaling and scale linking. International Adult Literacy Survey Technical Report. Ottawa, CA: Statistics Canada.