Kapitel 20
ICCS als repräsentative Schulleistungsstudie

Daniel Deimel & Nina Johanna Welsandt

20.1 Einleitung

Erläuterung Large-Scale-Assessment

Die International Civic and Citizenship Education Study (ICCS) versteht sich als *internationales Large-Scale-Assessment* (ILSA) und somit als Untersuchung, welche domänenspezifisch kognitive Kompetenzen erhebt, die für gesellschaftliche Teilhabe als erforderlich erachtet werden. Die Befunde können psychometrischen Standards entsprechend für eine definierte Kohorte in größeren Gebietseinheiten repräsentativ abgebildet und zufallskritisch abgesichert werden (Baumert et al., 2019). Neben der Erfassung der kognitiven Kompetenzen (vgl. Kap. 3) wird die Erfassung nicht vorrangig kognitiver Ergebnisse in ILSAs als gleichrangig betrachtet (Levin, 2013). Zusätzlich bildet ICCS als ILSA Kontexte ab, welche die relevanten Betrachtungsgrößen beeinflussen oder mit denen sie in Wechselwirkung stehen (Kuger et al., 2016).

Kapitelaufbau

In diesem Kapitel werden die Methoden erläutert, mit denen die repräsentativen Stichproben von ICCS konstruiert wurden. Weiter werden die Verfahren umrissen, die zur Datenaufbereitung und -skalierung genutzt wurden. Methodische Herausforderungen, die sich aus den genutzten Verfahren ergeben, und die für ICCS 2022 gewählten Lösungsansätze werden vorgestellt. ICCS 2022 wurde als direkte Nachfolgestudie zu ICCS 2009 und ICCS 2016 konzipiert. Aus Deutschland beteiligten sich mit Nordrhein-Westfalen (NRW) und Schleswig-Holstein (SH) zwei Bildungssysteme mit jeweils repräsentativen Erhebungen an ICCS 2022. Als Studie unter Gesamtverantwortung der International Association for the Evaluation of Educational Achievement (IEA) beachtet ICCS 2022 etablierte technische Standards für diese Art von Studien (Martin et al., 1999).

20.2 Stichprobe

Zielpopulation von ICCS 2022

Repräsentativität in der Sozialforschung bedeutet, dass nicht nur Aussagen über die spezifische Stichprobe der Studie getroffen werden können, sondern reliable (zuverlässige) Schätzungen für die Population im Sinne der Gesamtheit einer Gruppe gemacht werden. Die Zielpopulation der Schüler*innen von ICCS 2022 umfasst Schüler*innen der 8. Jahrgangsstufe, sofern dies die Stufe ist, die von 13.5 Jahre alten Schüler*innen zum Erhebungszeitpunkt durchschnittlich am häufigsten besucht wird. Ergänzt wird diese Stichprobe durch Lehrkräfte, welche zum Erhebungszeitraum in dem entsprechenden Jahrgang unterrichten, sowie durch Schulleitungen der teilnehmenden Schulen.

Ziel von ICCS ist es, mithilfe der Stichprobe möglichst genaue und repräsentative Aussagen über die Zielpopulationen der teilnehmenden Länder bzw. Bil-

Deimel, D. & Welsandt, N. J. (2024). ICCS als repräsentative Schulleistungsstudie. In H. J. Abs, K. Hahn-Laudenberg, D. Deimel & J. F. Ziemes (Hrsg.), *ICCS 2022. Schulische Sozialisation und politische Bildung von 14-Jährigen im internationalen Vergleich* (S. 377–391). Waxmann. CC BY 4.0. https://doi.org/10.31244/9783830998228.20

dungssysteme machen zu können. Hierzu werden verschiedene Verfahren des Stichprobendesigns angewendet, die im Folgenden kurz umrissen werden. Darauf aufbauend werden die geplanten und realisierten Stichproben für Schüler*innen, Lehrkräfte und Schulleitungen in den beiden aus Deutschland an ICCS 2022 beteiligten Bildungssysteme vorgestellt: Nordrhein-Westfalen (NRW) und Schleswig-Holstein (SH).

20.2.1 ICCS 2022 – teilnehmende Bildungssysteme

ICCS 2022 wurde in insgesamt 24 Bildungssystemen durchgeführt. Tabelle 20.1 führt die teilnehmenden Bildungssysteme und ausgewählte makrosystemische Rahmenbedingungen auf. **Fett** hervorgehoben sind die Länder und Regionen, die neben NRW und SH am europäischen Modul teilgenommen haben. In diesen 20 Bildungssystemen wurde aufgrund geografischer und politischer Nähe zusätzlich zu den allgemeinen Erhebungsinstrumenten der Europäische Schülerfragebogen eingesetzt. Diese Länder, beziehungsweise Regionen, sowie Serbien werden in diesem Bericht als „Vergleichsgruppe Europa“ (VG Europa) bezeichnet. Die europäische Vergleichsperspektive wird in diesem nationalen Bericht ausgewählt, weil bei aller Heterogenität die Kontexte der einbezogenen Schulsysteme in einem politisch und kulturell stärker vergleichbaren und (etwa über den Europarat oder der EU) verbundenen Rahmen verortet sind als die weiteren drei in ICCS 2022 repräsentierten internationalen Teilnahmeländer.

Erläuterung der Vergleichsgruppe Europa

In Tabelle 20.1 *kursiv* aufgeführt sind die Teilnahmeländer, die nicht bei der internationalen Kalibrierung von Skalen berücksichtigt wurden. Dies gilt einerseits für Benchmark-Teilnehmer sowie für Länder, deren Teilnahmequoten auf Ebene der Schulen und/oder Schüler*innen nicht die internationalen Vorgaben erreichen. Benchmark-Teilnehmer (vgl. Anmerkung 1) sind in dieser IEA-Studie ausschließlich NRW und SH. Das bedeutet, dass die beiden Bundesländer kein eigenständiges Mitglied der IEA sind, sondern eine abgrenzbare Teilregion innerhalb eines IEA-Mitgliedslandes (Deutschland). Anmerkung 2 bezieht sich auf das Erreichen der angestrebten Stichprobe und die Anmerkungen 3 bis 5 beziehen sich auf Aspekte, welche aufgrund des Stichprobendesigns einer besonderen Aufmerksamkeit in der Dateninterpretation für die jeweiligen Länder bedürfen.

NRW und SH als Benchmark-Teilnehmer

Tabelle 20.1: Teilnahmeländer ICCS 2022 inklusive makrosystemischer Kontextvariablen

Länder	Index der menschl. Entwicklung (HDI)	Bruttoinlandsprodukt (BIP) (USD pro Kopf)	Wahlbeteiligung bei den letzten nationalen Wahlen (%)	Frauenanteil im nationalen Parlament (%)	Liberal Democracy Index (V-Dem)	Regimetyp	Korruptionswahrnehmungsindex
Brasilien[2]	0.75	7 697	79.1[D]	14.8	0.53	ED	38
Bulgarien	0.80	12 222	39.3	24.2	0.61	ED	42
Dänemark[2]	0.95	68 008	84.2	43.6	0.89	LD	88
Deutschland (NRW)[1]	0.94[A]	51 204[A]	76.4[B]	35.0[A]	0.81[A]	LD[A]	80[A]
Deutschland (SH)[1,2]	0.94[A]	51 204[A]	78.2[B]	35.0[A]	0.81[A]	LD[A]	80[A]
Estland	0.89	27 944	63.7	29.7	0.85	LD	74
Frankreich	0.90	43 659	47.5	37.3	0.80	LD-	71
Italien	0.90	35 770	63.7	34.5	0.77	LD-	56
Kolumbien	0.75	6 183	55.0	29.0	0.55	ED	39
Kroatien[4]	0.86	17 748	46.9	23.2	0.65	ED+	47
Lettland[4]	0.86	21 080	59.4	30.0	0.74	LD	59
Litauen	0.88	23 737	47.8	27.0	0.73	ED+	61
Malta	0.92	34 218	85.6	5.1	0.64	ED+	54
Niederlande[3]	0.94	57 708	78.7	39.3	0.80	LD	82
Norwegen[4,5]	0.96	90 655	77.2	45.0	0.86	LD-	85
Polen	0.88	18 000	61.7	28.7	0.43	ED	56
Rumänien	0.82	14 927	32.0	18.4	0.55	ED	45
Schweden[4]	0.95	61 143	84.2	46.1	0.87	LD	85
Serbien	0.80	9 230	58.6	38.8	0.27	EA	38
Slowakei	0.85	21 783	65.8	21.3	0.78	LD-	52
Slowenien	0.92	29 291	71.0	40.0	0.71	ED+	57
Spanien	0.91	30 104	65.4	39.4	0.79	LD	61
Taiwan	0.93	33 190	74.9[C]	41.6	0.73	LD	68
Zypern	0.90	31 552	65.7	14.3	0.64	LD-	53

Fett hervorgehoben sind Bildungssysteme, die am europäischen Modul teilgenommen haben.
Kursiv gesetzt sind Länder, die für die Kalibrierung von Skalen nicht berücksichtigt wurden.

[1] Benchmark-Teilnehmer. Nicht Teil der VG Europa.
[2] Teilnahmequoten auf Schul- und/oder Schülerebene liegen unterhalb der internationalen Vorgaben, wodurch Repräsentativität nicht in gleicher Weise abgesichert ist.
[3] Vorgaben für Teilnahmequoten auf Schul- und/oder Schülerebene unter Einbezug von Ersatzschulen werden beinahe erreicht.
[4] Ausschlüsse der nationalen Zielpopulation liegen zwischen 5–10 %.
[5] Abweichende Zielpopulation: 9. Klasse.

Daten und folgende Anmerkungen sind, wenn nicht anders angegeben, aus dem Internationalen Bericht übernommen (Schulz et al., 2023b):
[A] Die Zahlen beziehen sich auf ganz Deutschland.
[B] Wahlbeteiligung bei der letzten Bundestagswahl 2021 auf Landesebene (eigene Recherche: IT.NRW bzw. Statistikamt Nord)
[C] Wahlpflicht
[D] Wahlpflicht, aber optional für Wähler*innen zwischen 16 und 18 sowie über 70
Quelle HDI: Human Development Report (UNDP, 2022)
Quelle BIP pro Kopf: World Bank Indicators, https://data.worldbank.org/indicator
https://www.idea.int/data-tools/data/voter-turnout; https://www.idea.int/data-tools/data/gender-quotas/country-view/290/35; https://www.idea.int/data-tools/data/voter-turnout/compulsory-voting

Daten des Demokratieindex und Regimetyps: https://www.v-dem.net/documents/29/V-dem_democracyreport2023_lowres.pdf
Regimetyp: LD – Liberal Democracy; ED – Electoral Democracy; EA – Electoral Autocracy; CA – Closed Autocracy
– deutet darauf hin, dass das Land unter Berücksichtigung von Unsicherheit auch zur niedrigeren Kategorie gehören könnte;
+ zeigt an, dass das Land möglicherweise auch zur höheren Kategorie gehört
Daten Korruptionswahrnehmungsindex: Transparency International https://www.transparency.org/en/cpi/2021
Quellen Taiwan: HDI – https://ws.dgbas.gov.tw/Download.ashx?u=LzAwMS9VcGxvYWQvMC9yZWxmaWxlLzExMDIwLzIyOTU5MS9iNDdhNmYyYy1jNjY2LTRjZDAtYmQ2Ni03OGEyYjMwMmM4MzkucGRm&n=TjExMTEwMTQucGRm&icon=.pdf
BIP pro Kopf – https://www.imf.org/en/Publications/WEO/weo-database/2023/April/weo-report?c=156,&s=NGDP_RPCH,NGDPD,PPPGDP,NGDPRPPPPC,NGDPDPC,PPPPC,&sy=2015&ey=2026&ssm=0&scsm=1&scc=0&ssd=1&ssc=0&sic=0&sort=country&ds=.&br=1

 © ICCS 2022

20.2.2 Stichprobendesign

Der Ansatz von ICCS 2022 (wie in vielen anderen ILSAs) ist es, Schüler*innen als Teil ihrer Klasse bzw. Lerngruppe repräsentativ für ihre Schule zu erfassen. Zusätzlich ist es aufgrund der Komplexität von Schulrealitäten, welche sich etwa durch ein mehrgliedriges Bildungssystem und eine ungleiche Verteilung von Schüler*innen mit internationaler Geschichte (zur Definition vgl. Kap. 18) konkretisiert, nicht hinreichend, eine einfache Zufallsstichprobe zu ziehen. Um diese Komplexität zu berücksichtigen und die Genauigkeit der Parameterschätzungen zu erhöhen, entspricht das Stichprobendesign einer stratifizierten zweistufigen Cluster-Stichprobe (Bortz & Döring, 2016, S. 310–317). Die Stichprobe wurde auf Grundlage gemeinsamer Planungen mit dem nationalen Studienzentrum durch die Sampling-Unit der IEA Hamburg gezogen.

Genutzte Stratifizierungsverfahren in ICCS 2022

In einem ersten Schritt wurde die Zielpopulation der Schülerschaft in mehrere Untereinheiten eingeteilt (*Stratifizierung*). In der *expliziten Stratifizierung* wurden Schulen mit ähnlichen Merkmalen zu sinnvollen Einheiten (Strata) zusammengefasst. Die *Strata* ergeben sich in NRW und SH zunächst durch die Schulformen der gegliederten Schulsysteme. Als weiteres Stratum wurde der Anteil der Schüler*innen mit Migrationshintergrund herangezogen; hier wurde der *cut-off*-Wert von 33 % innerhalb der Schülerschaft gewählt.[1] In anderen Ländern wurden häufig andere Strata implementiert, wie beispielsweise Regionen innerhalb eines Landes (vgl. Schulz et al., 2018). Ziel der Stratifizierung ist es, die Verteilung der Schüler*innen in der Zielpopulation auf die verschiedenen Schultypen innerhalb der Stichprobe abzubilden. Innerhalb dieser Strata erfolgte in einem zweiten Schritt eine Zufallsauswahl von Schulen unter Berücksichtigung der Anzahl der Schüler*innen im jeweiligen Jahrgang (*probability proportional to size*, *PPS sampling*). Dies soll ausgleichen, dass Schüler*innen an großen Schulen eine geringere Wahrscheinlichkeit haben, Teil der Stichprobe zu sein, als Schüler*innen an kleinen Schulen. Bei Gymnasien in SH sowie bei den Förderschulen in SH und NRW wurde die explizite Stratifizierung nach Migrationshintergrund aufgrund der geringen Größe von manchen sich daraus ergebenen Strata durch eine implizite Stratifizierung ersetzt, in der neben Schulgröße auch Migrationsanteil laut Schulstatistik bei der Sortierung der Schulen für die Ziehung nach PPS-Verfahren berücksichtigt wurde.

Oversampling zur Präzisionserhöhung einzelner Teilstichproben

Um Effekte, die in Zusammenhang mit der internationalen Geschichte von Schüler*innen stehen können, mit höherer Genauigkeit schätzen zu können, wurde zudem ein *Oversampling* von Schulen mit einem hohen Anteil an Schüler*innen mit internationaler Geschichte vorgenommen. Von diesen Schulen wurden also mehr erhoben, als eine rein proportionale Stichprobe vorgesehen hätte. Die beschriebenen Verfahren werden im Prozess der Gewichtung so berücksichtigt, dass Aussagen für die Population repräsentativ sind.

In einem letzten Schritt der Stichprobenziehung erfolgte eine zufällige Auswahl einer 8. Klasse innerhalb jeder der ausgewählten Schulen. In dieser Klasse werden alle Schüler*innen erhoben, welche keine Ausschlusskriterien erfüllen. An Förderschulen in NRW und SH wurden zudem, aufgrund der geringeren Anzahl von Schüler*innen pro Klasse, alle Klassen eines Jahrgangs zu *Pseudo-Klassen* zusammengefasst und für alle weiteren Verfahren wie die Schüler*innen derselben Klasse behandelt.

---

1 Maßgeblich sind hier die spezifischen Definitionen der beiden Bundesländer im Rahmen ihrer Schulstatistik, zur vertieften Auseinandersetzung siehe Kapitel 18.

Angestrebte Stichprobengröße

Aufgrund der Clusterung der Schüler*innen in Schulen und Schulklassen, sowie der daran anschließenden genutzten Verfahren der Stichprobenziehung, ist eine größere Stichprobe erforderlich, um eine ähnliche statistische Aussagekraft wie eine einfache Zufallsstichprobe geringerer Größe zu erreichen. Die angestrebte Präzision des Samples soll derselben Präzision einer hypothetischen einfachen Zufallsstichprobe (*effective sample size*) von 400 Schüler*innen entsprechen; mit einem 95 % Konfidenzintervall von ±0.1 Standardabweichungen (Martin et al., 1999). Übertragen auf die Metriken von ICCS bedeutet dies, dass der Standardfehler nicht mehr als 5 Skalenpunkte für den Wissenstest bzw. 0.5 Punkte für die Skalenwerte der Fragebogenmodule übersteigen soll. Als angestrebter Wert für IEA-Studien hat sich hierfür eine intendierte Stichprobengröße von 150 Schulen pro Bildungssystem etabliert, da diese in der Regel in einer realisierten Stichprobe von mindestens 3.000 Schüler*innen mündet.

Intendierte und realisierte Stichprobe

In jeder empirischen Studie ist davon auszugehen, dass die *realisierte Stichprobe* der *intendierten Stichprobe* nicht vollständig entspricht. Verweigert eine Schule ihre Teilnahme, kann hierfür in einer hinsichtlich der Größe ähnlichen Ersatzschule desselben Stratums eine zufällig ausgewählte Klasse erhoben werden. Steht eine Schule überhaupt nicht zur Verfügung (etwa aufgrund einer Schulschließung oder weil keine Klasse im entsprechenden Jahrgang vorhanden ist), wird sie nicht ersetzt. Während die öffentlichen Schulen in NRW aus Gründen der schulischen Qualitätssicherung zur Teilnahme verpflichtet waren, stand es Schulen in nicht öffentlicher Trägerschaft frei. In SH war die Teilnahme an ICCS 2022 für alle Schulen freiwillig.

Bestimmung der Güte einer realisierten Stichprobe

Für ICCS 2022 wurden internationale Vorgaben definiert, inwieweit die realisierte Stichprobe von der intendierten Stichprobe abweichen darf. Diese internationalen Vorgaben betreffen die Quote der realisierten Stichprobe bezüglich der gezogenen Stichprobe. In Bezug auf die Erhebung der Schüler*innen müssen mindestens 85 % der Schulen und 85 % der Schüler*innen erreicht werden. Alternativ darf das Produkt aus der Erfassungsquote von Schulen und Schüler*innen 75 % nicht unterschreiten. Länder, die diese Anforderungen erfüllen, werden der *Kategorie 1* zugeordnet, etwa NRW. Für diese Länder wird davon ausgegangen, dass Verzerrungen, die durch Unterschiede von realisierter und intendierter Stichprobe entstehen, vernachlässigbar klein sind.

Länder, die diese Anforderungen lediglich unter Berücksichtigung der Ersatzschulen erreichen und in denen mindestens 50 % der Schulen und Schüler*innen der ursprünglich gezogenen Stichprobe (also ohne Berücksichtigung der Ersatzschulen) erreicht werden, wie die Niederlande, werden *Kategorie 2* zugeordnet. Hier wird eine gewisse Verzerrung befürchtet, jedoch erwartet, dass diese durch den Einsatz von Gewichten ausreichend ausgeglichen werden kann.

Liegt die kombinierte Schul- und Schülerbeteiligungsquote trotz regelkonformer Umsetzung des Erhebungsprozesses auch nach Einbeziehung der Ersatzschulen bei unter 75 %, wird auf internationaler Perspektive eine eingeschränkte Vergleichbarkeit angenommen und entsprechend gekennzeichnet (z. B. Dänemark und Brasilien). Die Ergebnisse in diesen Ländern werden bei der Kalibrierung von Skalen nicht berücksichtigt und werden der *Kategorie 3* zugeordnet. SH ist in dieser Kategorie eingeordnet, wäre als Benchmark-Teilnehmer aber auch bei höherer Ausschöpfung nicht in den Kalibrierungsprozessen berücksichtigt worden.

Freiwilligkeit und Verpflichtung der Teilnahme für Schüler*innen

Die Teilnahmeentscheidung für individuelle Schüler*innen erfolgte in den verschiedenen Schulsystemen unterschiedlich. Für die Teilnahme an dem internationalen sowie dem europäischen Schülerfragebogen war in NRW und SH jeweils eine Elterngenehmigung erforderlich; zusätzlich konnten auch Schüler*innen mit

dieser Genehmigung die Teilnahme verweigern. In SH galten diese Bedingungen auch für den kognitiven Test, während dieser in NRW als Maßnahme der schulischen Qualitätssicherung verpflichtend gemacht wurde. Im internationalen Datensatz werden nur Klassen berücksichtig, bei denen die Teilnahmequote auf Schülerebene über 50 % lag, weil dort ansonsten systematische Verzerrungseffekte zu befürchten sind. Der Grund hierfür besteht darin, dass die Verweigerung der Teilnahme, etwa bei fehlenden Elterngenehmigungen, möglicherweise nicht zufällig, sondern systematisch ist. Der Ausschluss hängt möglicherweise mit Merkmalen der Schüler*innen zusammen, die ebenfalls Einfluss auf die Bearbeitung des Fragebogens oder auf die im Fragebogen erfragten Inhalte haben.

Darüber hinaus bestehen internationale Vorgaben darüber, unter welchen Bedingungen und in welchem Ausmaß Schulen oder Klassen bereits im Vorfeld der Zufallsauswahl ausgeschlossen werden dürfen. Auf Schulebene dürfen spezielle Schulformen (wie Förderschulen) ausgeschlossen werden, soweit sie sich curricular und unterrichtsmethodisch von Regelschulen bedeutsam unterscheiden. Dies betrifft in Deutschland die Förderschwerpunkte *Geistige Entwicklung, Hören & Kommunikation, Sehen, Körperliche Entwicklung* sowie *Schulen für Kranke*. Auf Klassenebene sind weitere Ausschlussgründe international definiert: Sofern eine Klasse ausschließlich aus Schüler*innen besteht, die aus körperlichen, emotionalen oder kognitiven Gründen nicht befähigt sind, an der Studie teilzunehmen, oder ausschließlich aus Schüler*innen besteht, die die Sprache der Erhebungsinstrumente seit weniger als einem Jahr sprechen, wird diese Klasse vor der Zufallsauswahl ausgeschlossen. Innerhalb der ausgewählten Klassen wurden Schüler*innen ausgeschlossen, denen eine Teilnahme aufgrund der vorgenannten Merkmale nicht möglich war. Auch testökonomische Aspekte spielen hier eine Rolle, da für eine relativ geringe Anzahl von Schüler*innen ein hohes Ausmaß von Anpassungen erforderlich wäre, ohne dass die Vergleichbarkeit der Materialien hinreichend empirisch abgesichert werden kann. Dies muss bei der Interpretation der Daten berücksichtigt werden.

Die internationalen Vorgaben legen fest, dass die Ausschlussquote 5 % der Schüler*innen für eine uneingeschränkte Vergleichbarkeit nicht übersteigen darf. In ICCS 2022 weisen Kroatien, Lettland, Norwegen und Schweden eine höhere Ausschlussquote von 5 bis10 % auf, was hinsichtlich der Vergleichbarkeit bei der Ergebnisinterpretation berücksichtigt werden muss. In NRW liegt die Ausschlussquote bei 3,4 %, in SH bei 2,1 %. Der Unterschied ist durch den Ausschluss von anthroposophisch ausgerichteten Schulen erklärbar: In NRW bildeten diese aufgrund der Anzahl ein eigenes Stratum. Als Schulen in privater Trägerschaft war ihnen die Teilnahme freigestellt. Es entschieden sich so viele Schulen gegen eine Teilnahme, dass sich die Teilpopulation in diesem Stratum nicht mehr adäquat abbilden lies und das gesamte Stratum als „ausgeschlossen" behandelt werden musste. In SH ist die Anzahl entsprechender Schulen so klein, dass sie im Stratum „integrierte Gesamtschule" eingeschlossen waren.

Eingeschränkte Vergleichbarkeit zu ICCS 2016

NRW hat als einziges Bundesland in Deutschland bereits an ICCS 2016 teilgenommen, jedoch unterscheidet sich die Stichprobe von der in ICCS 2022 in NRW: Damals lag die Ausschlussquote insgesamt bei 7 %, u. a. wurden alle Förderschulen unabhängig vom Förderschwerpunkt ausgeschlossen. Weiter zeigte die intendierte Stichprobe im internationalen Vergleich mit 35 % eine wesentlich geringe Ausschöpfung und somit weniger Präzision der Populationsschätzungen. Aufgrund der darin resultierenden geringeren Vergleichbarkeit, insbesondere im unteren Leistungsbereich (Ziemes et al., 2017), erscheint es als wenig sinnvoll, Unterschiede im Zeitvergleich (ICCS 2022 zu ICCS 2016) statistisch systematisch

auswerten und absichern zu wollen. Gleichwohl wird an einzelnen Stellen in diesem Bericht auf Ergebnisse von ICCS 2016 verwiesen, wenn sich grundlegende oder deutliche Unterschiede zeigen. Zeitreihenvergleiche in anderen Schulsystemen sind im internationalen Bericht dokumentiert (Schulz et al., 2023b).

20.2.3 Intendierte und realisierte Stichproben in NRW und SH

Die intendierte und realisierte Stichprobe von Schüler*innen, Lehrkräfte und Schulleitung in NRW und SH findet sich in Tabelle 20.2. Im Folgenden wird diese zusammenfassend dargestellt sowie die Aussagekraft der jeweiligen Stichproben erläutert. Schulen gelten als „teilnehmend" für die Schülerstichprobe, wenn wenigstens zu 50 % der Schülerschaft in der gezogenen Klasse Daten des kognitiven Tests oder der Fragebogenmodule vorliegen, bzw. für die Stichprobe der Lehrkräfte, wenn mindestens für 50 % der Lehrkräfte in Jahrgangsstufe 8 gültige Daten vorliegen. Sollte diese Schwelle unterschritten werden, wird die jeweilige Schule für die Stichprobe der Schüler*innen bzw. Lehrkräfte so behandelt, als hätte sie nicht an ICCS 2022 teilgenommen, d. h., eventuell vorliegende Daten werden von weiteren Analysen ausgeschlossen.

Kriterien für den Einschluss von Fällen in den Datensatz

Tabelle 20.2: Stichproben und geschätzte Population für NRW und SH

	Nordrhein-Westfalen			Schleswig-Holstein		
	intendiert	realisiert	geschätzte Population	intendiert	realisiert	geschätzte Population
Schüler*innen	3 998	3 269	166 438	2 312	1 488	27 485
Lehrkräfte in Jahrgang 8[1,2]	-	2 916	46 349	-	173	-
Lehrkräfte in Jahrgang 8 (Sozialwissenschaften)[1,2,3]	-	1 012	16 230	-	53	-
Schulleitungen[4]	152	141	1 945	121	75	491
Schulen (gültige Schülerdaten)[5]	152	145	1 875	121	84	376
Schulen (gültige Lehrerdaten)[5]	152	129	1 875	121	13	-

1 Listung der Anzahl der Lehrkräfte der Zielpopulation erfolgt erst durch die Schulen und ist nicht Teil der Stichprobenplanung.
2 Aufgrund unzureichender Stichprobenausschöpfung der Lehrkräfte in SH erfolgt für diese keine Schätzung der Population.
3 Filterfrage: „Unterrichten Sie in der Jahrgangsstufe 8 ein Fach, das im weiteren Sinne mit politischer Bildung und Demokratieerziehung zu tun hat"
4 Die Samplingprozesse streben eine repräsentative Stichprobe für Schüler*innen an. Gewichtete Schätzungen der Population der Schulleitungen sind daher von größerer Ungenauigkeit betroffen.
5 Die realisierte Stichprobe bezieht sich auf Schulen mit mindestens 50 % Teilnahme auf Ebene der Schüler*innen bzw. Lehrkräfte.

IEA: International Civic and Citizenship Education Study

© ICCS 2022

Die Stichprobe der Schüler*innen in NRW wurde zufriedenstellend ausgeschöpft. Insgesamt 145 Schulen zählen als „teilnehmend", für 3 269 Schüler*innen liegen gültige Daten vor. Dies entspricht einer kombinierten Erfassungsquote von 85 % und damit der Kategorie 1. Sie ist ohne Einschränkung geeignet, repräsentative Aussagen über Schüler*innen der 8. Jahrgangsstufe der entsprechenden Schulformen zu treffen. Eine schlechtere Ausschöpfung war aufgrund der freiwilligen Teilnahme für Schulen in SH erwartbar. In SH liegen für insgesamt 84 Schulen Daten von 1 488 Schüler*innen vor. Die Schülerstichprobe SH erreicht damit eine Erfassungsquote von 61 % und wurde in Kategorie 3 eingeordnet. Das heißt, dass mit ihr unter Berücksichtigung der geringeren Ausschöpfung repräsentative Aussagen

Aussagekraft der Schülerstichproben in NRW und SH

über die Zielpopulation möglich sind, diese aber nicht in der gleichen Weise abgesichert werden können.

Vertiefende Betrachtung der realisierten Stichproben SH

Erste Analysen zeigen, dass alle Strata in SH durch die realisierte Stichprobe abgedeckt wurden. Förderschulen, denen allerdings nur ein sehr kleiner Anteil (geschätzt 1,2 %) der Schüler*innen in SH zuzuordnen sind, haben vergleichsweise seltener teilgenommen (45 % Ausschöpfung der Stichprobe), während an Gymnasien die intendierte Stichprobe mit 95 % fast vollständig ausgeschöpft wurde. In den beiden Strata der integrierten Gesamtschulen wurde mit 66 % (Schulen mit < 33 % Anteil Schüler*innen mit Migrationshintergrund) bzw. 70 % (Schulen mit > 33 % Anteil Schüler*innen mit Migrationshintergrund) der größere Teil der Schülerschaft erreicht. Insgesamt könnte dies darauf verweisen, dass insbesondere bei nicht gymnasialen Schulformen in SH aufgrund der weniger zur Verfügung stehenden Schülerdaten nur etwas ungenauere Schätzungen vorgenommen werden können, was sich in höheren Standardfehlern (siehe Abschnitt 20.2.4) widerspiegelt. Tabelle 20.3 gibt ergänzend den Anteil der Schüler*innen sowie die geschätzte Population in den verschiedenen Referenzkategorien für Geschlecht und internationale Geschichte für NRW und SH an.

Aussagekraft der Lehrerstichproben in NRW und SH

Die Stichprobe der Lehrkräfte in NRW ist geeignet, um repräsentative Aussagen über Lehrkräfte, die in Jahrgangsstufe 8 unterrichten, zu treffen, wenn auch mit kleineren Einschränkungen. Die Stichprobe umfasst 2.916 Lehrkräfte an 129 Schulen und wurde mit einer kombinierten Erfassungsquote von 71 % der Kategorie 3 zugeordnet. Hier zeigt sich hinsichtlich der Ausschöpfungslücken nur wenig Systematik: Während beispielsweise in beiden gymnasialen Strata, im Stratum Hauptschule mit einem Anteil von über 33 % Schüler*innen mit Migrationshintergrund sowie im Stratum Förderschulen eine nahezu vollständige Teilnahme vorliegt (jeweils 90–100 %), zählen in beiden Strata der Sekundarschulen sowie der Realschulen mit einem Anteil von über 33 % Schüler*innen mit Migrationshintergrund etwa 71 % bis 75 % der Schulen als teilnehmend. Auch dies kann im Vergleich zu höheren Standardfehlern führen. An der Lehrkräftebefragung in SH haben insgesamt 580 Lehrkräfte teilgenommen, allerdings wurde nur an 13 Schulen (mit 173 teilgenommenen Lehrkräften) eine Teilnahmequote von über 50 % erreicht. Die erhobenen Daten können damit lediglich als große Gelegenheitsstichprobe gelten, aber keinen Anspruch auf Repräsentativität erheben. Daher haben wir uns entschieden, in diesem Bericht keine Daten für Lehrkräfte in SH in Tabellen zu berichten.

Ausschöpfung der Stichproben für Schulleitungen in NRW und SH

Für die Befragung der Schulleitungen war jeweils ein Mitglied der Schulleitung dazu aufgefordert, den Fragebogen zu bearbeiten. Die intendierte Stichprobe wurde in NRW – wo die Teilnahme verpflichtend war – fast vollständig (93 %) ausgeschöpft. In SH beteiligten sich – unter der Bedingung einer freiwilligen Teilnahme – immerhin 62 % der Schulleitungen. Es ist zu beachten, dass die Stichprobenplanung vor allem eine Repräsentativität für Schüler*innen anstrebt und alle Schätzungen für Schulleitungen durch eine größere Unsicherheit und größere Standardfehler äußert.

Nationale Zusatzstichprobe Lehrkräfte

In NRW und SH wurde weiter eine ergänzende Lehrerstichprobe definiert. Diese nationale Zusatzstichprobe umfasst Lehrkräfte in sozialwissenschaftlichen Fächern, die nicht in Jahrgang 8 unterrichtet haben und somit nicht Teil der internationalen Stichprobe sind. Da der Fokus in diesem Bericht auf dem internationalen Vergleich liegt, erfolgt keine Darstellung von Ergebnissen der Zusatzstichprobe in diesem Band.

Tabelle 20.3: Schüler*innen der achten Jahrgangsstufe in NRW und SH

	Schüler*innen Nordrhein-Westfalen			Schüler*innen Schleswig-Holstein		
	N[1]	Est[2]	%[3]	N[1]	Est[2]	%[3]
Geschlecht						
Weiblich	1 434	74 607	45	678	12 318	45
Männlich	1 776	88 917	54	764	14 325	52
Weiteres	53	2 616	2	44	825	3
Internationale Geschichte[4]						
Min. ein Elternteil in DE geboren	1 954	100 145	67	1 203	22 258	84
Schüler*in in DE geboren, Eltern zugewandert	619	32 870	22	147	2 493	9
Schüler*in und Eltern zugewandert	319	15 606	11	94	1 893	7

1 Absolute Fallzahl.
2 Gewichteter Anteil der Schüler*innen an der Population.
3 Anteil an der gewichteten Population in Prozent.
4 Zur Definition in diesem Bericht siehe Kapitel 18.

IEA: International Civic and Citizenship Education Study

© ICCS 2022

20.2.4 Gewichtung und Schätzung von Stichproben und Messfehlern

Gewichtung als Verfahren zur Absicherung von Repräsentativität

Durch die IEA-Sampling-Unit wurden den Datensätzen Gewichte hinzugefügt. Bei einer Gewichtung werden für statistische Auswertungen Antworten von Schüler*innen in dem Maße stärker berücksichtigt, in dem sie in der realisierten Stichprobe im Vergleich zur intendierten Stichprobe unterrepräsentiert sind. Hierdurch können anhand der Daten eher Aussagen über die Population als nur die spezifische Stichprobe gemacht werden. Eine Gewichtung verbessert dadurch auch die internationale Vergleichbarkeit von Daten, da hier Effekte von unterschiedlich realisierten Stichproben – begrenzt – ausgeglichen werden können. Für jede Klasse wurde ein individueller Gewichtungsfaktor anhand eines Basisgewichts (inverse Selektionswahrscheinlichkeit separat für jeden Schritt der Stichprobenziehung) und eines Adjustierungsfaktors generiert, welcher anhand der tatsächlich teilgenommenen Schüler*innen aus allen Selektionsstrata berechnet wurde. Im Gewichtungsprozess wurde angenommen, dass fehlende Werte sich über die berücksichtigten Strata hinaus nicht systematisch erklären lassen (*missing completely at random*).

Schätzung von Standardfehlern

Durch Gewichtungen, stratifizierte Stichprobenziehung und Unterschiede zwischen der intendierten und realisierten Stichprobe droht eine Unterschätzung der Varianzen und damit der Standardfehler. Einige statistische Verfahren versuchen, das Maß der Verzerrung anhand der realisierten Stichprobe zu schätzen und realistischere Standardfehler zu präsentieren: Die *Jackknife*-Methode nutzt die Selektionsstrata, die eingangs beschrieben worden sind. Innerhalb der expliziten Strata werden Schulen in Paaren angeordnet, die sich ähneln im Hinblick auf vordefinierte Merkmale wie etwa Schulform und Anteil an Schüler*innen mit Migrationshintergrund. Im ICCS-2022-Datensatz zu NRW entstanden so 75 *Jackknife*-Zonen, in SH 43 Zonen. Zur Schätzung des Standardfehlers wird jede Berechnung 75 Mal (bzw. 43 Mal) durchgeführt. In jeder Durchführung der Berechnung wird eine Schule ausgeschlossen und die gepaarte Schule dafür doppelt gewichtet. Aus der Varianz der Ergebnisse werden die Standardfehler der Populationsschätzung berechnet.

Zusammenhang zwischen Ausschöpfung von Stichproben, Gewichtung und Standardfehlern

Umso mehr die realisierte Stichprobe von der intendierten abweicht, desto intensiver muss dies im Gewichtungsprozess berücksichtigt werden (durch höhere Gewichte) und desto größer wird der Standardfehler. Der größere Standardfehler bedeutet eine größere Ungenauigkeit der Schätzung und führt weiterhin dazu, dass Unterschiede zwischen den Ergebnissen von (Teil-)Gruppen größer sein müssen, um statistisch signifikant zu werden. Wenn in ICCS 2022 international höhere Standards für die Zuverlässigkeit der Schätzung gelten, als in einzelnen Teilstichproben erreicht werden, wird in jeder Ergebnistabelle dieses Berichts darauf hingewiesen, wenn die Repräsentativität der Ergebnisse nicht in der gleichen Weise abgesichert ist. Weiter werden für geschätzte Mittelwerte stets die Standardfehler dieser Mittelwerte berichtet. Der Standardfehler (*SE*) ist ein Maß für die Genauigkeit einer Schätzung und wird insbesondere bedeutsam, wenn etwa die Signifikanzen von Gruppenunterschieden berechnet werden sollen. Mit einer 95 % Wahrscheinlichkeit liegt der wahre Wert einer Messung (also hier der tatsächliche Wert in der Population) innerhalb des Intervalls von 1.96 Standardfehlern über oder unter dem berichteten Wert. Ein Unterschied wird als statistisch signifikant bezeichnet, wenn es sehr wahrscheinlich ist, dass sich dieser von einem vorher festgelegten Wert, beispielsweise Null oder dem Wert einer Vergleichsgruppe, unterscheidet. In der Regel wird in den Sozialwissenschaften eine maximale Irrtumswahrscheinlichkeit von 5 % als Richtwert genommen. Dies bedeutet, dass es eine Wahrscheinlichkeit von weniger als 5 % gibt, dass das beobachtete Ergebnis in der Stichprobe zufällig auftritt, wenn kein tatsächlicher Effekt vorliegt.

Um Prozesse der Gewichtung zum Ausgleich der heterogenen Abdeckung einzelner Strata zu rechtfertigen, wird davon ausgegangen, dass die vergleichsweise geringere Beteiligung in einzelnen Strata nicht auf einer systematischen und inhaltlich bedeutsamen Beeinflussung beruht. Inhaltlich relevant und systematisch wäre etwa, wenn an bestimmten Schulen innerhalb nur eines Stratums deutlich mehr finanzielle Mittel zur Verfügung stehen, die dazu führen, dass diese Schulen eher teilnehmen und Schüler*innen gleichzeitig besser vorbereitet sind. Für ICCS 2022 liegen weder in NRW noch in SH Hinweise auf eine solche Beeinflussung der Teilnahmeentscheidung von Schulen vor.

20.3 Datenauswertung und Skalierung

Soweit nicht anders dargestellt entsprechen die Methoden der Skalierung, Gewichtung, und Schätzung der Standardfehler den in ICCS 2016 genutzten Methoden (Schulz et al., 2018). Vertiefende Informationen werden internationalen technischen Bericht zu ICCS 2022 vorgestellt, dessen Erscheinen für Mitte 2024 erwartet wird. In den folgenden Abschnitten werden Konzepte vorgestellt, die für ein Verständnis der quantitativ-empirischen Berechnungen in diesem Bericht unmittelbar relevant sind. Für die Berechnungen aller Ergebnisse wurde der IDB-Analyzer Version 5 der IEA (IEA, 2023) in Verbindung mit R Version 4.3.1 (R Core Team, 2023) sowie SPSS Version 29 (IBM, 2022) genutzt. Der IDB Analyzer ist eine frei zugängliche Analysesoftware, die Makros und einen Syntaxgenerator zur Verfügung stellt, mit denen notwendige Verfahren zur Berechnung von Schätzungen und Standardfehlern korrekt angewendet werden. Grundlage aller hier berichteten Ergebnisse ist die Datenversion, die von IEA Hamburg am 14.9.2023 für die Berichtslegung freigegeben wurde.

20.3.1 Umgang mit fehlenden Daten

Verschiedene Kategorien fehlender Werte in ICCS

Fehlende Werte erhöhen die Unsicherheit beim Schätzen von Testergebnissen oder anderen Skalenwerten für (Teil-)Populationen. Im Wissenstest wurden unterschiedliche Formen von fehlenden Werten unterschieden: Eine fehlende Antwort entsteht etwa, wenn eine Aufgabe nicht Teil des Testmaterials der konkreten Person war, beispielsweise, weil die Aufgabe in einem anderen *Booklet* ist. Weiter entsteht ein ausgelassenes Item, wenn Schüler*innen eine Aufgabe übersehen oder bewusst ausgelassen haben. Für den Wissenstest existierte die zusätzliche Kategorie „nicht erreicht", wenn Schüler*innen Aufgaben am Ende des Tests aus Zeitmangel nicht bearbeiten konnten. Für unterschiedliche Bereiche der Fragebögen gab es schließlich die Kategorie „logisch nicht zutreffend": Wenn sich beispielsweise bestimmte Fragen nur an Lehrkräfte mit einem bestimmten Unterrichtsfach richteten, erhielten andere Lehrkräfte für die betreffenden Fragen diesen Wert.

In diesem Bericht wurden fehlende Werte in Hintergrundvariablen, wie Geschlecht, Alter oder sozioökonomischer Status der Eltern, nicht imputiert, also nicht anhand anderer vorhandener Werte geschätzt und ersetzt. Dies führt dazu, dass bei Berechnungen, die Unterschiede zweier Schülergruppen hinsichtlich eines bestimmten Merkmals betrachten, die Schüler*innen nicht berücksichtigt werden, die keine Angaben zu dieser Hintergrundvariable gemacht haben (Ausschluss *listwise*).

20.3.2 Skalierungsmodelle

Vorteile von Skalierungsverfahren

Die Überführung von Antworten zu einzelnen Items in eine gemeinsame Skala hat mehrere Vorteile: Erstens wird vermieden, dass ein sprachlich oder kulturell bedingt anderes Verständnis oder ein Lesefehler bei einem einzelnen Item sich in gleicher Weise in der Bewertung des Konstrukts widerspiegelt, wie dies bei einem Einzelitem der Fall wäre. Zweitens werden Konstrukte erfassbar, die abstrakter sind, als es die Formulierung konkreter Items zulässt. Durch Skalenwerte wird eine genauere Schätzung von Konstrukten ermöglicht, als dies durch einzelne Fragen möglich wäre. Neben der Berechnung eines Summen- oder Mittelwertes gibt es auch andere Methoden, die statistisch aufwendiger, jedoch gerade in Schulleistungsstudien verbreitet sind (van de Vijver et al., 2019), da sie präzisere Schätzungen für die Verteilung von Merkmalen in der Population bieten. Dazu wird unter anderem berechnet, wie gut die einzelnen Items das Konstrukt repräsentieren, für das sie eingesetzt wurden, oder wie wahrscheinlich eine bestimmte Antwort auf ein Item ist, wenn die Information zu den anderen Items einer Skala vorliegen. Solche Berechnungen könnten dazu genutzt werden, auch bei Vorliegen fehlender Werte genaue Schätzungen für die Population zu erzielen.

Nutzung von IRT-Verfahren in ICCS

In ICCS 2022 werden zur Skalierung Verfahren der Item-Response-Theorie (*IRT*) genutzt. Die Grundannahme ist, dass sich für jede Person ein Fähigkeitswert schätzen lässt und für jede Aufgabe, beziehungsweise jedes Item, ein Schwierigkeitswert. Je schwerer ein Item ist, desto seltener wird die richtige Antwort (Leistungstests) oder eine bestimmte Antwortoption (Einstellungsfragebogen) gewählt. Nach diesem Modell beeinflusst die Fähigkeit von Personen die Wahrscheinlichkeit, ein Item auf eine bestimmte Art zu bearbeiten. Die IRT erlaubt es, Schüler*innen und Items auf der gleichen Skala anzuordnen (vgl. auch Abbildung 3.5 in Kapitel 3). Vertiefend zur IRT siehe auch Moosbrugger (2012). Für ILSAs wie ICCS 2022 basiert die Verwendung von IRT-Modellen auf der Annah-

me, dass die Schwierigkeitsrangfolge der Wissensitems zueinander über Gruppen hinweg relativ einheitlich ist (Äquidistanz). Inwieweit diese Annahme erfüllt wird, lässt sich durch komplexe Verfahren überprüfen. Je ähnlicher die Abstände zwischen den Aufgaben zwischen verschiedenen Teilgruppen (z. B. nach Schulsystem, Land oder Geschlecht) sind, desto eher kann davon ausgegangen werden, dass das gleiche Merkmal, wie etwa politisches Wissen, mit den gleichen Items in gleicher Weise gemessen werden kann.

Im Unterschied zur Berechnung eines arithmetischen Mittels ergibt sich die Metrik einer Skala nicht aus der Anzahl der Antwortoptionen eines Items. Der Mittelwert und eine Standardabweichung können beliebig festgelegt werden. Unabhängig vom festgelegten Mittelwert und der Standardabweichung bleiben die Relationen zwischen den Werten der Länder, Klassen und Schüler*innen erhalten.

Metrik von ICCS 2022

In ICCS ist die Skala des Leistungstests für den internationalen Vergleich auf einen Mittelwert von 500 und eine Standardabweichung 100 skaliert, für Ergebnisse der Fragebögen auf einen Mittelwert von 50 mit einer Standardabweichung von 10. Die Werte sind weiter jeweils an den Werten von ICCS 2009 bzw. ICCS 2016 kalibriert, um eine Vergleichbarkeit über die Studienzyklen hinweg zu ermöglichen.

Kalibrierung des Wissenstests und der Skalen der ICCS-2022-Fragebögen

Die Kalibrierung des Wissenstests von 2022 (und zuvor ICCS 2016) erfolgte an den Werten von ICCS 2009. Gruppenmittelwerte von ICCS 2022 können im Verhältnis zur Vergleichsgruppe 2009 sowie 2016 interpretiert werden. Weiterhin können Länder, die 2009 und/oder 2016 schon teilgenommen haben, reliabel Veränderungen im politischen Wissensbestand ihrer 14-jährigen Schüler*innen einschätzen (*Multikohortenvergleich*). Eine Kalibrierung erfolgte weiter für Einstellungsskalen, die schon in ICCS 2009 und/oder 2016 eingesetzt wurden. Die Mittelwerte und die Standardabweichungen wurden in Hinblick auf die Werte von 2009 (bei unveränderten Skalen seit 2009) bzw. 2016 (bei neuen oder veränderten Skalen in 2016) kalibriert. Somit können Abweichungen vom Skalenwert 50 für ICCS 2022 als Abweichungen vom Gesamtmittel von 2009 und/oder 2016 interpretiert werden. Ob eine Skala hiervon betroffen ist, wird im jeweiligen Kapitel berichtet. Aufgrund der Stichprobengüte von NRW in ICCS 2016, das als einziges deutsches Bundesland teilgenommen hatte, ist ein Zeitvergleich nicht in gleicher Form möglich (siehe 20.2.1). Genaue Angaben zur Analyse und Kalibrierung des Wissenstests und der weiteren Skalen werden ausführlich im internationalen technischen Bericht dargelegt.

Die internationale Nutzung von IRT Modellen ist herausfordernd, da länderspezifische Abweichungen einerseits durch ein fehlerhaftes oder unvollständiges Modell der Fähigkeit, andererseits durch kulturelle, sprachliche oder curriculare Unterschiede verursacht werden können (Schulz & Fraillon, 2011). Die Analyse dieser Abweichungen von Modellen stellt einen komplexen Forschungsbereich für sich da und ist Teil der Qualitätssicherungsprozesse der Studie (vgl. Kap. 19 zu Invarianz und DIF).

20.3.3 Schätzung des politischen Wissens und latenter Konstrukte des Einstellungsfragebogens

Für die Erfassung politischen Wissens in ICCS 2022 wurden alle Items aus dem Wissenstest in insgesamt vierzehn Clustern zusammengefasst. Jede*r Schüler*in erhielt eines von 14 sogenannten *Booklets*. In jedem *Booklet* waren drei Cluster mit Testaufgaben sequentiell angeordnet; dieses Testdesign ist in Tabelle 20.4 ab-

gebildet. Gekennzeichnet sind hier die Testcluster, die lediglich in Ländern mit computerbasierter Testung eingesetzt wurden (siehe Kap. 19). Jede*r Schüler*in bearbeitete so nur einen Teil des gesamten Tests von ICCS 2022, nämlich 3/14. Dieses Booklet-Rotationsdesign hat den Vorteil, dass mögliche Reihungseffekte statistisch kontrolliert werden können. Zusätzlich kann eine größere Zahl an Aufgaben eingesetzt werden, wodurch der Fähigkeitsraum, in dem die Schüler*innen eingeordnet werden, präziser geschätzt werden kann, ohne dass sich die Bearbeitungslänge des Tests erhöht. Aufgrund der unterschiedlichen *Booklets* wird die Schätzung für den Vergleich von Populationen und für hinlänglich große Teilpopulationen angemessener.

Rotationsdesign von ICCS 2022

Tabelle 20.4: Booklet-Design des kognitiven Tests

Booklet Nummer	alle Booklets	Position 1	Position 2	Position 3
1	Einleitung, Anleitung, Übungsfragen	C1	C12*	C9
2		C2	C13*	C10
3		C3	C14*	C11
4		C4	C1	C12*
5		C5	C2	C13*
6		C6	C3	C14*
7		C7	C4	C1
8		C8	C5	C2
9		C9	C6	C3
10		C10	C7	C4
11		C11	C8	C5
12		C12*	C9	C6
13		C13*	C10	C7
14		C14*	C11	C8

*Module mit Computer-Enhanced Items (CEI) (vgl. Kap. 3.2.2)

IEA: International Civic and Citizenship Education Study © ICCS 2022

Zur Schätzung des politischen Wissens wurde ein *Plausible-Value*-Ansatz verwendet (Kaplan & Su, 2016). Für die Berechnung von plausiblen Werten wird anhand der bearbeiteten Aufgaben und Hintergrundvariablen zunächst für jede*n Schüler*in eine Verteilung des politischen Wissens bzw. die auf alle Items bezogene Fähigkeit geschätzt (vgl. Schulz et al., 2018, S. 133). Somit kann auf Grundlage des Antwortverhaltens in den drei der 14 bearbeiteten Clustern die auf alle Items bezogene Fähigkeit geschätzt werden. Anschließend wurden aus dieser Verteilung für jede*n Schüler*in fünf plausible Werte (*Plausible Values*) zufällig ausgewählt, die für alle Berechnungen, die politisches Wissen beinhalten, genutzt werden. Durch die fünffache Wiederholung der Analysen und die Zusammenführung der Ergebnisse kann die Unsicherheit der Schätzung mit einbezogen und die Präzision der Populationsschätzung insgesamt erhöht werden, gleichzeitig sind sie zur Individualdiagnostik einzelner Schüler*innen nicht geeignet. Die Nutzung plausibler Werte wird mit der *Jackknife*-Methode (siehe 20.2.4) verbunden. Die Berechnungen werden damit bis zu 375 Mal durchgeführt, um einen möglichst genauen Schätzer des Standardfehlers der Populationsschätzung zu erhalten.

Schätzung des politischen Wissens in ICCS 2022

Umfang und Aufbau des ICCS-2022-Wissenstests

Der Wissenstest besteht aus insgesamt 141 Items. Von diesen wurden 121 in den elf Clustern angeordnet, die sowohl in papier- als auch computerbasierten Varianten vorhanden waren. Die 20 übrigen Items wurden als *Computer-Enhanced-Items* (siehe Kap. 3) ausschließlich computerbasiert erhoben. In NRW und SH wurden ausschließlich computerbasierte Testungen durchgeführt (siehe Kap. 19). Ein größerer Teil des Testmaterials wurde auf Grundlage angepasster Inhaltsbereiche für ICCS 2022 neu entwickelt (Schulz et al., 2023a). Zur Absicherung der Vergleichbarkeit der Tests zu ICCS 2009 und ICCS 2016 wurde ein Teil der Aufgaben aus vorherigen Zyklen übernommen: Insgesamt 55; 26 davon wurden in 2009 und 2016 eingesetzt; 29 Aufgaben wurden erstmalig in ICCS 2016 genutzt.

Antwortformate der ICCS-2022-Fragebögen

Antworten im Wissenstest werden zumeist dichotom interpretiert; sie werden entweder als richtig oder falsch gewertet (einzelne Aufgaben weichen hiervon ab und vergeben bis zu zwei Punkte, siehe Kap. 3.2.1). Einstellungsfragen hingegen haben keine richtigen oder falschen Antworten. Antworten können nur im Hinblick auf ein bestimmtes Kriterium mehr oder weniger wünschenswert sein. In der Regel wurden in ICCS 2022 sogenannte Likert-Formate eingesetzt. Diese bestehen aus einer Aussage und einer balancierten Antwortskala, von der eine Antwortmöglichkeit gewählt werden kann. Balanciert bedeutet, dass es gleichviele zustimmende wie ablehnende Antwortmöglichkeiten gibt. In den Sozialwissenschaften gilt die Annahme, dass es bei Likert-Formaten wenig problematisch ist, die Antworten für die Auswertung in einen numerischen Wert zu überführen.

Skalierung mit Items der ICCS-2022-Fragebögen

Über polytome Raschanalysen (*Partial-Credit*-Modelle) wird zusätzlich eine präzisere Einschätzung über die Stärke der Einstellung angestrebt, indem die Annahme berücksichtigt wird, dass für die Wahl der nächst stärkeren Antwort nicht immer das gleiche Maß an stärkerer Zustimmung notwendig ist (Schulz, 2017). Beispielsweise vermeiden bisweilen Menschen (teilweise kulturabhängig) extreme Antworten. Für die Skalen wurde also nicht nur jedes Item auf einer Skala angeordnet, sondern alle Antwortoptionen jedes Items. Dem gleichen Prinzip des Wissenstests folgend können nun Merkmalsausprägungen der Schüler*innen, also Einstellungen, auf der gleichen Skala angeordnet werden. Eine höhere Ausprägung der Einstellung geht dann einher mit einer höheren Wahrscheinlichkeit, die stärker zustimmende Antwortoption zu wählen. Die Übergangspunkte zwischen zwei Antwortabstufungen, die mit einem bestimmten Maß von Einstellungen korrespondieren, sind die sogenannten Thurstone-Grenzwerte. Ebenfalls kann die Einstellung relativ gut geschätzt werden, selbst wenn nicht jede einzelne Einstellungsfrage beantwortet wurde. Die Skalenwerte basieren auf den Weighted Likelihood Estimate (*WLE*) Scores, die aus den Antwortmustern und Itemparametern (Schwierigkeit) berechnet werden.

Fragebogenitems und Skalen als Manifestierung latenter Konstrukte

Als Grundlage für die Durchführung der IRT-Analysen wurden für die Einstellungsitems konfirmatorische Faktorenanalysen durchgeführt. Konfirmatorische Faktorenanalysen überprüfen, ob die Items eine gemeinsame Varianz teilen. Wenn dies im ausreichenden Maß der Fall ist, so lässt sich begründen, dass sich alle Items auf ein gemeinsames dahinterliegendes latentes Konstrukt beziehen und darum in einer Skala zusammengeführt werden können. Diese Analysen werden ausführlich im technischen Report beschrieben. Sie geben jedoch keine Information darüber, ob sich das Zusammenfassen der Items auch theoretisch begründen lässt. Die theoretische Basis für den Wissenstest sowie Einstellungsskalen wird jeweils im *Assessment Framework* kurz hergeleitet (Schulz et al., 2023a). Weiter werden die Konstrukte auch in den jeweiligen Kapiteln dieses Berichts in einen ausführlicheren Theoriekontext eingebettet. Zusätzlich geben wir im Bericht jeweils die interne Konsistenz einer Skala an. Diese wird mit *Cronbach's Alpha*

numerisch abgebildet, wobei Werte nahe 1 eine hohe Reliabilität anzeigen und darauf hinweisen, dass die Items einer Skala gut miteinander korrelieren und ein einheitliches Konstrukt messen. Zunächst wurde die interne Konsistenz der Skalen in den jeweiligen Teilnahmeländern berechnet, wir berichten den Mittelwert aller Länder der Kategorie 1 und 2, die keine Benchmarking-Teilnehmer sind.

Interne Konsistenz der Skalen in ICCS 2022

Literatur

Baumert, J., Maaz, K., Lühe, J. & Schulz, S. (2019). Bildungsungleichheit und Bildungsarmut – Der Beitrag von Large-Scale-Assessments. In Quenzel (Hrsg.), *Handbuch Bildungsarmut* (S. 261–285). Springer VS. https://doi.org/10.1007/978-3-658-19573-1_10

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

IBM. (2022). *IBM SPSS statistics for Windows, Version 29.0*. IBM Corp.

IEA. (2023). *IEA IDB Analyzer (Version 5.0)*. IEA.

Kaplan, D. & Su, D. (2016). On matrix dampling and iImputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, *41*(1), 57–80. https://doi.org/10.3102/1076998615622221

Kuger, S., Klieme, E., Jude, N. & Kaplan, D. (Hrsg.). (2016). *Assessing contexts of learning: An international perspective*. Springer. https://doi.org/10.1007/978-3-319-45357-6

Levin, H. M. (2013). The utility and need for incorporating noncognitive skills into large-scale educational assessments. In M. von Davier & E. Gonzalez (Hrsg.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (S. 67–86). Springer. https://doi.org/10.1007/978-94-007-4629-9_5

Martin, M. O., Rust, K. & Adams, R. J. (1999). *Technical standards for IEA studies*. IEA.

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227–274). Springer. https://doi.org/10.1007/978-3-642-20072-4_10

R Core Team. (2023). *R: A language and environment for statistical computing: Version 4.3.1*. https://www.R-project.org

Schulz, W. (2017). Scaling of questionnaire data in international large-scale assessments. In R. D. Adams (Hrsg.), *Implementation of large-scale education assessments* (S. 384–410). John Wiley. https://doi.org/10.1002/9781118762462.ch15

Schulz, W., Carstens, R., Losito, B. & Fraillon, J. (Hrsg.). (2018). *ICCS 2016 technical report*. IEA.

Schulz, W. & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation*, *17*(6), 447–464. https://doi.org/10.1080/13803611.2011.630559

Schulz, W., Fraillon, J., Losito, B., Agrusti, G., Ainley, J., Damiani, V. & Friedman, T. (2023a). *IEA International Civic and Citizenship Education Study 2022 assessment framework*. Springer. https://doi.org/10.1007/978-3-031-20113-4

Schulz, W., Fraillon, J., Losito, B., Agrusti, G., Damiani, V., Ainley, J. & Friedman, T. (2023b). *Education for Citizenship in Times of Global Challenge: IEA International Civic and Citizenship Education Study 2022 International Report*. IEA.

UNDP. (2022). *Human Development Report 2021-22*. United Nations Development Programme. http://report.hdr.undp.org

van de Vijver, F. J. R., Jude, N. & Kuger, S. (2019). Challenges in International LargeScale Educational Surveys. In L. Suter, E. Smith & B. D. Denman (Hrsg.), *The SAGE handbook of comparative studies in education* (S. 83–102). Sage. https://doi.org/10.4135/9781526470379.n6

Ziemes, J. F., Jasper, J., Deimel, D., Hahn-Laudenberg, K. & Abs, H. J. (2017). ICCS 2016. Design und methodisches Vorgehen. In H. J. Abs & K. Hahn-Laudenberg (Hrsg.), *Das politische Mindset von 14-Jährigen: Ergebnisse der International Civic and Citizenship Education Study 2016* (S. 47–76). Waxmann.

Bildnachweis S. 375:
Fotogruppe Die Bilderkammer Max-Brauer-Schule, Frau mit Fisch, Deutscher Jugendfotopreis 2014/DHM