Kapitel 19

ICCS als international vergleichende Studie

Entwicklung der Instrumente und Durchführung der Erhebung

Johanna F. Ziemes & Beatriz Matafora

19.1 Einleitung und Organisationsstruktur

International vergleichende Schulleistungsstudien, oft auch *International Large Scale Assessments* (ILSAs) genannt, sind Studien, die bestimmte Altersgruppen oder Klassenstufen im Hinblick auf Leistung oder andere Merkmale in verschiedenen Bildungssystemen untersuchen und vergleichen (vgl. Johansson, 2016). ILSAs haben verschiedene Potenziale für Politiker*innen, Praktiker*innen und Forscher*innen. Zuvorderst erlauben sie durch die gewonnenen Daten und durchgeführten Analysen, Annahmen über das eigene Bildungssystem in Frage zu stellen. Zum Beispiel zeigen Studien wie ICCS (Hahn-Laudenberg & Abs, 2017), TIMSS (Stubbe et al., 2020) und PISA (Weis et al., 2019), dass in den Teilnahmeländern deutliche Leistungsunterschiede zwischen Schüler*innen mit niedrigerem und höherem sozioökonomischen Status bestehen. Die Größe des Unterschiedes variiert aber deutlich zwischen den Systemen. Durch den internationalen Vergleich bietet sich die Möglichkeit, Einblicke in die Variabilität von Bildungsprozessen zu gewinnen. Dabei ermöglichen ILSAs die Identifikation von *Best Practices* aus Bildungssystemen. Bei der Interpretation der Analysen und vor einem Transfer muss berücksichtigt werden, dass sich (gesetzliche) Regelungen oder Lehrstrategien nur selten unbesehen auf andere Bildungssysteme übertragen lassen. Vielmehr müssen die institutionellen und kulturellen Begebenheiten und Besonderheiten der jeweiligen Bildungssysteme berücksichtigt werden (Johansson, 2016; Niemann, 2015).

Potenziale international vergleichender Schulleistungsstudien

Der Transfer von Praktiken ist an das Wissen um kulturelle und politische Gegebenheiten gebunden …

Um ihre Funktion zu erfüllen, sind ILSAs angewiesen auf die Berücksichtigung anspruchsvoller Erhebungs- und Auswertungsmethoden. Nationale und regionale Besonderheiten sind auch für die Durchführung von ILSAs bedeutsam. Diesen Besonderheiten als Herausforderungen produktiv zu begegnen, legt die Grundlage für die internationale Vergleichbarkeit und die Möglichkeit des Transfers. Dieses Methodenkapitel bietet einen Einblick in Maßnahmen, welche in der Vorbereitung der Erhebungsinstrumente und Durchführung der Erhebung in ICCS 2022 getroffen wurden, um eine möglichst hohe internationale Vergleichbarkeit der Ergebnisse sicherzustellen. Eine allgemeinere Darstellung von Maßnahmen, die die internationale Vergleichbarkeit von Bildungsstudien verbessern, kann an anderer Stelle nachgelesen werden (z. B. Rutkowski & Rutkowski, 2019). Zunächst werden die beteiligten Organisationen vorgestellt. Es folgt ein Unterkapitel über die Adaptation und Übersetzung der Erhebungsinstrumente (19.2) und abschließend ein Unterkapitel, indem die Durchführung der Erhebung und ihrer Qualitätssicherung beschrieben werden (19.3).

… und stellt hohe Anforderungen an die Methoden der Studien.

ICCS 2022 versteht sich als der dritte Zyklus der *International Civic and Citizenship Education Study* (nach ICCS 2009 und ICCS 2016). ICCS ist die Nachfolgestudie der *Civic Education Study* 1999 (CivEd). Alle Studien wurden von

Ein Netzwerk von Institutionen war für die Studiendurchführung verantwortlich.

Ziemes, J. F. & Matafora, B. (2024). ICCS als international vergleichende Studie. Entwicklung der Instrumente und Durchführung der Erhebung. 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. 365–374). Waxmann. CC BY 4.0. https://doi.org/10.31244/9783830998228.19

der *International Association for the Evaluation of Educational Achievement* (IEA) initiiert. Die IEA hat für ICCS 2022 den *Australian Council for Educational Research* (ACER) in Kooperation mit dem *Laboratorio di Pedagogia Sperimentale* (ROMA TRE) als internationales Studienzentrum beauftragt. Diese Organisationen waren für die nähere Konzeption der Studie und die Vorbereitung der Befragungsinstrumente verantwortlich. In der konkreten Ausgestaltung arbeiteten sie eng mit den Forschungszentren der Teilnahmeländer zusammen.

Neuerung für den Zyklus ICCS 2022: Computerbasierte Erhebung in vielen Ländern.

Im Unterschied zu ICCS 2016 wurde im ICCS 2022er Zyklus die Erhebung in den meisten teilnehmenden Bildungssystemen computerbasiert durchgeführt. ACER beauftragte das Unternehmen RM mit der Erstellung und Verwaltung der digitalen Übersetzungsplattform, sowie der Programmierung des digitalen *Players*, also der Software, die an Schulen zur Erhebung mittels Computern eingesetzt wurde. Für den Prozess der Adaptation, Übersetzung, Erhebung, Codierung und Feldarbeiten wurden von ACER Anleitungen zur Verfügung gestellt, die bei Bedarf übersetzt worden sind. Diese wurden zum Teil von IEA Hamburg und IEA Amsterdam mitverfasst.

Die IEA unterstützt bei der Qualitätskontrolle aller Prozesse.

Die IEA ist in Teilorganisationen gegliedert, die in unterschiedlicher Funktion in ICCS 2022 involviert waren: Die Mitarbeiter*innen von IEA Amsterdam waren etwa verantwortlich für die Überprüfung der Übersetzungen, die Qualitätssicherung auf internationaler Ebene, sowie für die Veröffentlichung und weiteren Verbreitung des internationalen Studienberichts. Mitarbeitende aus IEA Hamburg haben die Stichprobenplanung, Datenverarbeitung und Verwaltung durchgeführt. Auch Richtlinien zur Qualitätsüberwachung und der Feldarbeit wurden hier entwickelt.

Nationale Studienzentrum unterstützen Fragebogenkonstruktion und führen Datenerhebung durch.

Die nationalen Studienzentren der teilnehmenden Bildungssysteme hatten bei der Entwicklung der Studie eine eigene Rolle: Sie haben die Entwicklung der Fragebögen unterstützt und waren u. a. für die Übersetzung der Erhebungsmaterialien und die Erhebung in den Bildungssystemen verantwortlich. Darüber hinaus lieferten sie im Rahmen der nationalen Kontextbefragung Informationen hinsichtlich ihrer Bildungssysteme und dem Stand der politischen Bildung und Demokratieerziehung, sowohl im Allgemeinen, als auch speziell in Bezug auf die ICCS-Zielgruppe (Jahrgangsstufe 8, bzw. der Jahrgangsstufe, in der Schüler*innen mit 13,5 Jahren am ehesten sind). Das nationale Studienzentrum in Deutschland besteht aus einem Konsortium mehrerer Organisationen: Hierzu gehören die Abteilung *Educational Research and Schooling* der Universität Duisburg-Essen, die Abteilung Bildung und Demokratiepädagogik im Kontext von Migration und Integration an der Universität Leipzig, sowie das DIPF Leibniz-Institut für Bildungsforschung und Bildungsinformation in Frankfurt. Für die Durchführung der Erhebung hat das nationale Studienzentrum eine Abteilung von IEA Hamburg beauftragt, die sich auf die Durchführung von Feldarbeiten internationaler und nationaler *Large-Scale-Assessments* in Deutschland spezialisiert hat. IEA Hamburg übernahm damit für Nordrhein-Westfalen (NRW) und Schleswig-Holstein (SH) das Training der Testleiter*innen, den direkten Schulkontakt, die Organisation der Testungen sowie die Sammlung der Daten. Das nationale Studienzentrum hat hier eigene Beobachtungen zur Qualitätssicherung durchgeführt.

19.2 Adaptation und Übersetzung

Die Erhebungsinstrumente mit Wissenstest und Fragebögen waren die Grundlage für die Befragung in ICCS 2022. ILSAs können nur dann zu vergleichbaren Ergebnissen führen, wenn Konstrukte ausgewählt werden, die in verschiedenen Ländern und Kulturen ähnlich verankert sind (van de Vijver et al., 2019). Für ICCS stellt das Vorhandensein einer repräsentativen Demokratie einen wesentlichen gemeinsamen kulturellen Anker dar. Wenn Items innerhalb von Fragebatterien z. B. in einzelnen Ländern unterschiedlich funktionieren, wird dies als *different item functioning* (DIF), kurz DIF, bezeichnet. Dieses Phänomen kann mit kulturellen und regionalen Besonderheiten zusammenhängen (Cuellar et al., 2021; Klieme & Baumert, 2001). Auch Übersetzungs- und Adaptationsmethoden, die Auswahl der Übersetzer*innen, unübersetzbare sprachliche Eigenheiten und Inkongruenzen bei Abkürzungen können DIF auslösen (El Masri et al., 2016). DIF kann durch die Anwendung verschiedener Maßnahmen minimiert werden.

Zu untersuchende Konzepte müssen in den verschiedenen Ländern ähnlich verankert sein.

Fragebögen werden so konstruiert, dass sie in allen Ländern möglichst ähnlich funktionieren.

Zunächst brauchen Studien ein theoretisches Fundament: Die konzeptuelle Grundlage der Erhebungsinstrumente war das *Assessment Framework* (Schulz et al., 2023). In diesem wurden Konstrukte definiert, die erhoben werden sollten, und in einen theoretischen Zusammenhang eingeordnet. In der Entwicklung des Assessment Frameworks und der Instrumente wurde Feedback der aktuellen nationalen Forschungszentren berücksichtigt. Auch Rückmeldungen zu der letzten Studie konnten hier Eingang finden. Auf Grundlage der ausgewählten Konstrukte wurden Instrumente entwickelt, adaptiert und übersetzt. Dabei waren in manchen der Bildungssysteme Übersetzungen in mehrere Sprachen erforderlich; z. B. werden für polnisch- und russischsprachige Schulen in Litauen auch Erhebungsinstrumente in Polnisch und Russisch angeboten. Ziel der Adaptation und Übersetzung war es nicht, eine exakte Wort-für-Wort-Übersetzung zu erreichen, sondern eine angemessene Übertragung der Konzepte in die Zielsprache und den Zielkontext (Rutkowski & Rutkowski, 2019). Die Vorbereitung der Testinstrumente bestand aus mehreren Schritten, die teils nacheinander und teils gleichzeitig durchlaufen wurden. Der Prozess war hochstandardisiert, um sicherzustellen, dass die Fragen in allen Sprachen möglichst ähnliche Bedeutungen, Leseschwierigkeit und Itemschwierigkeiten aufwiesen. Zudem sollte er dazu beitragen, dass kulturelle Verzerrungen (biases) vermieden werden (Koršňáková et al., 2020).

Das Assessment Framework dient der inhaltlichen Verankerung der zu messenden Konzepte.

Übersetzungen orientieren sich stärker an der Bedeutung als am genauen Wortlaut.

Begriffe, die im Studienkontext besonders sensibel für kulturelle Unterschiede waren, wurden vor der eigentlichen Übersetzung adaptiert. Der Prozess der Adaptation geht über die einfache Übersetzung hinaus: Beispiele dafür sind Begriffe wie der Name des nationalen Parlaments oder die angemessenen Bezeichnungen für international anerkannte Bildungsabschlüsse (ISCED-Stufen). Für die Adaptation wurden die Erhebungsinstrumente auf Englisch zunächst als *National Adaptation Form* (NAF) mit den nationalen Studienzentren geteilt. Die zu adaptierenden Begriffe waren gekennzeichnet. Die nationalen Studienzentren achteten bei der Adaptation nicht nur auf die Wortbedeutung, sondern auch auf die kulturellen und politischen Besonderheiten des Erhebungslandes. Nachdem die Dokumente zurück an das ACER und an das internationale Studienzentrum geschickt wurden, wurden die vorgenommenen Adaptiationen überprüft und das nationale Studienzentrum erhielt Rückmeldungen. Diese Schritte konnten nach Bedarf mehrfach durchlaufen werden bis alle Beteiligten den Adaptationen zustimmten. Abbildung 19.1 verdeutlicht diesen Prozess anhand des Begriffs „*local government*“. Das Ziel bestand darin, Begriffe zu finden, die nicht nur international äquivalent sind, sondern auch von den Schüler*innen der beiden teilnehmenden

In der Adaptation werden Begriffe angemessen in den Zielkontext übertragen.

deutschen Bundesländer gut verstanden werden. Das zweite Beispiel beschreibt den Entwicklungsprozess der Formulierung eines Items nach Rückmeldungen aus dem Feldtest. Eine detailliertere Beschreibung des Feldtests erfolgt weiter unten.

In den Items des Wissenstests wurden auch Namen für fiktive Personen adaptiert. Das nationale Studienzentrum adaptierte in diesem Fall mit dem Ziel, verschiedene Namen einzusetzen, die bestmöglich die real vorhandene Diversität der Schülerschaft berücksichtigen. Hierzu führten Mitarbeitende eine Durchsicht von Namen in deutschen standesamtlichen Quellen und weiterer Listen u. a. arabischer, slawischer und romanischer Vornamen durch. Bevorzugt wurden Namen, die in verschiedenen sprachlichen und kulturellen Kontexten verwendet wurden. Zentral bei der Auswahl war die Zugänglichkeit der Namen für die Schüler*innen. Weiterhin wurden geschlechterneutrale Namen bevorzugt (z. B. Sammy). Die Namen wurden zufällig auf die Items verteilt, um die Gefahr kulturalistischer Zuschreibungen zu minimieren.

Die eigentliche Übersetzung fand in mehreren Schritten statt.

Die Übersetzung der Test- und Fragebogenmaterialien fand anschließend in mehreren Schritten statt: Das gesamte Material wurde unabhängig von mindestens zwei Mitgliedern des nationalen Forschungszentrums aus dem Englischen ins Deutsche übersetzt. Im Anschluss wurden die Übersetzungen miteinander verglichen. Zusätzlich fand, soweit die Fragen unverändert waren, ein Abgleich mit dem Material statt, welches für ICCS 2016 in NRW eingesetzt worden war. Die Instrumente von ICCS 2016 und 2022 haben, wie es für international vergleichende Studien üblich ist, große Überschneidungen. In Teamgesprächen wurde die beste Übersetzung ausgewählt oder auf Grundlage der verschiedenen Vorschläge eine neue entwickelt, welche die Stärken verschiedener Übersetzungen kombinierte. Die Übersetzungen wurden anschließend von IEA Amsterdam überprüft und kommentiert. Der Austausch mit externen Übersetzer*innen wird als *Translation Verification* bezeichnet. Wie im Falle der Adaptation, konnte es mehrere Feedbackschleifen geben. Die letzten Entscheidungen über die Übersetzungen wurden vom nationalen Studienzentrum getroffen. Die Übersetzung und das Feedback fanden in einem von RM entwickeltem Online-Tool statt.

Hier kontrollierten sich mehrere Personen und Institutionen gegenseitig.

Zusätzlich wurde ein Pilot und ein Feldtest durchgeführt, um die Instrumente zu erproben.

Die Erhebungsinstrumente wurden vor ihrer Verwendung in der Hauptstudie von Schüler*innen in einem Piloten und einem Feldtest erprobt. Der Pilot wurde in Deutschland vollständig vom nationalen Forschungszentrum durchgeführt. Nach beiden Durchgängen wurde das Fragebogenmaterial überarbeitet. Zum Teil gaben Schüler*innen explizite Rückmeldungen zu den Instrumenten, z. B. bei nicht verstandenen Begriffen. Nach dem Feldtest fanden auf internationaler Ebene Berechnungen zur Dimensionalität, zur internationalen Invarianz der Skalen, sowie dazu statt, inwiefern einzelne Aufgaben des Tests in einzelnen Ländern aus der ansonsten beobachteten Rangfolge der Schwierigkeit herausfielen (DIF). Die betroffenen Items wurden durch das nationale Studienzentrum überprüft und sofern möglich angepasst, um die internationale Vergleichbarkeit zu verbessern. Anderes Erhebungsmaterial wurde nach dem Feldtest aussortiert, weil es die erforderlichen Gütekriterien nicht hinreichend erfüllte.

Abbildung 19.1: Austausch zu Übersetzung und Adaptation

Beispiel für Veränderungen im Adaptationsprozess

Im englischen Fragebogen findet sich eine Frage zum Vertrauen in verschiedene Institutionen, Medien und Gruppen. Im Item "The [local government] of your town or city" wurde [local government] für ICCS 2016 mit „Stadtrat oder Gemeinderat" adaptiert. Für den Zyklus 2022 schlug das deutsche Übersetzungsteam zunächst „Verwaltung" vor, mit der Überlegung, dass Schüler*innen eher mit Verwaltungsstellen als den gewählten Gremien interagieren und die gewählten Gremien in NRW (ICCS intern = DNW) und SH (= DSH) unterschiedlich bezeichnet werden.

Sie kommentierte den Adaptationsvorschlag mit: *"Verwaltung" implies administration more than legislation. We are unsure which part of the administration is meant here. "Stadtparlament" is not the technical term but could be understood in DSH and DNW*.

Das internationale Studienzentrum antwortete: "*Again this is a hard question, we note that "Stadtrat oder Gemeinderat" was used in 2016, could this not be adapted here as well?"*

Das *"again"* bezieht sich hier auf einen Kommentar zu einem anderen Item, in welchem ebenfalls [local government] adaptiert werden musste. Zu dieser (hier nicht abgedruckten) Frage hatte das internationale Studienzentrum geschrieben: *"This is a hard question, we would recommend using the level that would most likely have the power to enact the changes described in the responses."*

Das deutsche Übersetzungsteam antwortete: *Concerning "Stadtverwaltung": The problem is that the same word has different legal meanings in DNW and DSH. "Stadtrat" (DSH) is the deputy mayor and head of the city administration. Stadtrat (DNW) or Gemeindevertretung (DSH) would be the legislation; Stadtverwaltung the administration in both constituent states. On which aspect of local government shall the item focus on? Following our understanding of local government, we would propose: Stadtverwaltung oder Gemeindeverwaltung.*

Und das international Studienzentrum stimmte dem Vorschlag zu: *"what you propose is fine as it sounds like this will cater for both DSH and DNW."*

Die finale Adaptation war schließlich „Deiner Stadtverwaltung oder Gemeindeverwaltung".

Beispiel für Veränderungsprozesse der internationalen Items durch Rückmeldung aus den Teilnahmeländern

Im Feldtest wurde in europäischem Fragebogen ein Item vorgeschlagen (ES4G12E):

- [EU] policies are influenced by the richest member states.
  (Die Politik der EU wird von den reichsten Mitgliedsstaaten beeinflusst.)

Aus den Teilnahmeländern wurde die Rückmeldung gegeben, dass das Item unpräzise formuliert sei, da formal alle Mitgliedsstaaten die EU Politik beeinflussen und die Aussage daher unabhängig von der Bewertung des relativen Einflusses in Bezug auf die reichsten Mitgliedstaaten zugestimmt werden müsste. Die überarbeitete Version macht in der Hauptstudie die Intention deutlicher:

- [EU] policies are too strongly influenced by the richest member states.
  Die Politik der EU wird zu stark von den reichsten Mitgliedsstaaten beeinflusst.

Nicht in allen Ländern wurde computerbasiert erhoben.

In diesem ICCS Zyklus stand das internationale Studienzentrum vor der zusätzlichen Herausforderung, die Vergleichbarkeit der Ergebnisse der zwei Erhebungsmodi (computerbasiert oder papierbasiert) sicherzustellen, sowohl im Vergleich mit 2016, als auch im Vergleich zwischen den Bildungssystemen. Die computergestützte Erfassung war im Zyklus ICCS 2022 eine Neuerung und einige Länder verzichteten auf die computerbasierte Erfassung und nutzten stattdessen Papierfragebögen. Die Gewährleistung der Vergleichbarkeit der Ergebnisse zwischen der papierbasierten und der computergestützten Befragung ist notwendig, wenn für Länder ein Trend oder wenn Unterschiede zwischen Ländern berichtet werden sollen, die teilweise computerbasiert und teilweise papierbasiert ihre Befragungen durchführten (Schulz et al., 2023). In Bildungssystemen mit einer hinlänglich

abgesicherten repräsentativen Teilnahme in ICCS 2016 und computerbasierter Erhebung im Jahr 2022 wurde daher zusätzlich zur regulären Erhebung eine *Bridging-Study* durchgeführt. In diesen Ländern wurden sowohl Papierinstrumente als auch Computer eingesetzt. Mit den so gewonnenen Daten wurden Analysen durchgeführt, um die Äquivalenz der Erhebungsformate abzusichern. Aufgrund der eingeschränkten Repräsentativität der Daten für NRW in ICCS 2016 wurde für die deutschen Bildungssysteme keine *Bridging-Study* durchgeführt.[1]

Analysen stellten die Vergleichbarkeit von papier- und computerbasierten Instrumenten fest.

Die *Layout Verification* folgte der Übersetzung und Adaptation und war gefolgt von der Überprüfung der Funktionalität des *Players* (bei computer-basierten Assessment, wie in NRW und SH) oder einer Überprüfung der Seitenformatierung (bei der papierbasierten Erhebung, wie in Dänemark). Im computerbasierten Modus waren alle Elemente, die nicht übersetzt oder verändert werden sollten, vor Änderungen geschützt. Übersetzer*innen konnten nur die Texte bearbeiten, aber keine Positionen oder Layoutelemente verändern (Koršňáková et al., 2020). In der *Layout Verification* wurden in mehreren Schleifen Fehler identifiziert und behoben, um eine möglichst fehlerfreie Präsentation der Befragung sicherzustellen. Beispielsweise wurde im Rahmen dieses Arbeitsschritts daraufhin gewirkt, dass die Antwortformate bei längeren Itembatterien auch dann sichtbar bleiben, wenn ein*e Teilnehmer*in nach unten scrollt. Durch Verzögerungen in dem Prozess musste der Beginn der Haupterhebung verschoben werden. Ein letztes Item mit einer verpixelten Grafik erhielt zwei Wochen nach dem Start der Haupterhebung ein Update, welches den Grafikfehler behob.

Die computerbasierte Erhebung erforderte eine genaue Überprüfung des Erhebungsprogrammes.

19.3 Beschreibung der Erhebung

ICCS 2022 setzt sich aus verschiedenen Modulen zusammen, an denen nicht nur Schüler*innen, sondern auch Lehrkräfte und Schulleitungen teilnehmen. Tabelle 20.1 stellt die Fragebögen dar, die im Rahmen der Studie eingesetzt wurden, mit Angaben zu ihrer Dauer, den Befragten und der Art der Durchführung (Schulz et al., 2023, S. 96).

Der Pilot dient der ersten Erprobung der Instrumente.

Der Pilot, in dem einige neu entwickelte Fragebögen zum ersten Mal getestet wurden, fand (eingeschränkt durch die Auswirkungen der Pandemie) von Februar bis Mail 2020 an fünf Schulen in NRW statt. Der Pilot wurde noch mit Papierinstrumenten mit einer Gelegenheitsstichprobe von Schüler*innen der Zielklasse (Jahrgangsstufe 8) durchgeführt. Mit dem Feedback aus den Daten und dem der Schüler*innen wurden die Fragebögen überarbeitet. Dann wurde vom April 2021 bis zum Juni 2021 der Feldtest durchgeführt. An ihm nahmen 23 Schulen in NRW und SH mit insgesamt 1,175 Schüler*innen teil. In Deutschland war die IEA Hamburg für die Durchführung der Erhebung verantwortlich, bei der entweder von IEA Hamburg vorbereitete Laptops oder Schulcomputer mit USB-Sticks verwendet wurden. Im Feldtest nahmen trotz pandemiebezogener Einschränkungen fast alle Bildungssysteme mit einer ausreichenden Anzahl von Schulen teil, sodass erste vergleichende Skalen- und Faktorenanalysen durchgeführt werden konnten, um die Qualität der Messinstrumente zu analysieren.

Im Feldtest werden die Instrumente bereits international statistisch verglichen.

1 Teilgenommen an der *Bridging-Study* haben: Bulgarien, Estland, Italien, Kolumbien, Litauen, Lettland, Malta, Norwegen, Schweden, Slowenien und Taiwan.

Auf Grundlage der Analysen wurde das Test- und Fragebogenmaterial verfeinert und reduziert. Die Haupterhebung fand in Deutschland vom 14.03.2022 bis zum 23.06.2022 statt, wobei bis zum 05.10.2022 vereinzelte Nacherhebungen an wenigen Schulen stattfanden, um eine hohe Teilnahmequote sicherzustellen. Nacherhebungen wurden durchgeführt, wenn bei einer ersten Erhebung weniger als 50% der Schüler*innen einer Klasse anwesend waren. Nur Klassen, bei denen mindestens 50% der Schüler*innen, welche diese Klasse besuchten, teilgenommen haben, wurden in den Datensatz mit aufgenommen.

Die Haupterhebung in NRW und SH fand im zweiten Schulhalbjahr statt.

Die teilnehmenden Schulen wurden zufällig ausgewählt (vgl. Kap. 20). Jede teilnehmende Schule bestimmte eine Schulkoordination. Schulkoordinator*innen waren zumeist Lehrkräfte, die als Kontaktperson an der Schule fungierten. Sie standen im Kontakt zu IEA Hamburg, planten die Erhebungszeiten, organisierten die Elterngenehmigungen und den Raum für die Erhebung an der Schule. Schließlich gaben sie auch die *Login*-Daten für die Lehrkraft- und Schulfragebögen an die jeweiligen Personen weiter. Bei der Bewältigung dieser Aufgaben wurden sie von einem Manual, IEA Hamburg und dem nationalen Studienzentrum unterstützt.

Lehrkräfte unterstützten die Organisation an den Schulen.

Dies gilt auch für die Befragung der Lehrkräfte und Schulleitungen.

Die Erhebungen selbst wurden nicht von den Schulkoordinator*innen, sondern von Testleiter*innen durchgeführt. Die Testleiter*innen wurden von IEA Hamburg für die Durchführung der Erhebung an den Schulen ausgewählt, trainiert und eingeteilt. In der Regel handelte es sich hierbei um Lehramtsstudierende oder andere Personen mit einem akademischen pädagogischen oder psychologischen Hintergrund. Zu ihren Aufgaben gehörte das Bereitstellen von Laptops oder die Vorbereitung von Schulcomputern für die Erhebung. Sie erklärten den Schüler*innen die Abläufe der Erhebung, glichen die Teilnahmelisten ab und sorgten dafür, dass die Schüler*innen die Fragebögen ohne Einfluss von Dritten innerhalb der vorgegebenen Zeiten beantworten konnten. Die Testleiter*innen folgten bei den Erhebungen strengen, international einheitlichen Vorgaben und sie dokumentierten besondere Ereignisse und Störungen.

Die Testleiter*innen waren trainierte Personen mit einem pädagogischem Hintergrund.

Tabelle 19.1: Übersicht über die ICCS 2022 Testinstrumente und Erhebung

Instrument	Durchführungsart	Dauer	Zielgruppe
Internationaler Wissenstest (StT)	**Computer**- oder Papierbasiert	genau 45 Minuten	Schüler*innen
Pause		5 Minuten	
Internationaler Schülerfragebogen (StQ)	**Computer**- oder Papierbasiert	40-50 Minuten	
Pause		10 Minuten	
Europäischer Fragebogen (StE)	**Computer**- oder Papierbasiert	ca. 20 Minuten	
Oder			
Lateinamerikanischer Fragebogen (StL)	Computer- oder Papierbasiert	ca. 10 Minuten	
Lehrkräftefragebogen (TcQ)	**Online** oder Papier	ca. 30 Minuten	Lehrkräfte
Schulfragebogen (ScQ)	**Online** oder Papier	ca. 30 Minuten	Schulleitung

Anmerkung: Fett markiert ist die in NRW und SH genutzte Durchführungsart; Schüler*innen wurden in folgenden Ländern mit Papierinstrumenten befragt: Dänemark, Kroatien, Niederlande, Polen, Serbien, Rumänien.

IEA: International Civic and Citizenship Education Study © ICCS 2022

Schüler*innen bearbeiteten einen Test und zwei Fragebögenteile.

Im Wissenstest wurden auch interaktive Items eingesetzt.

Die Befragung der Schüler*innen bestand in NRW und SH aus drei Teilen: Dem Wissenstest (StT), dem internationalen Fragebogen (StQ) und dem europäischen Fragebogen (StE). Der Wissenstest umfasste elf reguläre Cluster von Items mit *Single-Choice*-Fragen und offenen Items. In Ländern mit computerbasierter Erhebung wurden drei zusätzliche digitale Cluster eingesetzt. Digitale Cluster enthielten auch Items mit interaktiven Elementen, die in einer papierbasierten Erhebung so nicht realisiert werden könnten. (siehe Kap. 3). Es wurden also insgesamt 14 Cluster eingesetzt. Aus jeweils drei Clustern wurde in unterschiedlicher Zusammenstellung eines von 14 Testheften gebildet (rotiertes Booklet-Design, vgl. Kap. 20.3). Testhefte wurden Schüler*innen zufällig zur Bearbeitung zugewiesen. Nach dem kognitiven Test hatten Schüler*innen eine kurze Pause, bevor sie den internationalen Fragebogen bearbeiteten. Dieser enthielt Fragen zum Elternhaus der Schüler*innen, deren Erfahrungen u. a. in der Schule und Fragen zum politischem Mindset. Neben den kognitiven Dimensionen wurden auch Aspekte der Identität und Selbstwirksamkeit, der Partizipation und verschiedener politischer Einstellungen berücksichtigt.

Das europäische Modul wurde in 18 Bildungssystemen eingesetzt.

Der Europäische Fragebogen wurde in NRW, SH und den anderen 18 teilnehmenden europäischen Bildungssystemen eingesetzt. Ein lateinamerikanischer Fragebogen wurde in Brasilien und Kolumbien durchgeführt. Diese regionalen Module enthielten Fragen, denen eine besondere Relevanz in den jeweiligen Regionen beigemessen wurde. Das europäische Modul umfasste z. B. Fragen zur europäischen Identität der Schüler*innen und zu europäischen Institutionen. In SH war die Erhebung mit dem europäischen Modul abgeschlossen. In NRW wurde nach der ICCS 2022 Befragung eine Pause von 20 Minuten eingelegt und im Anschluss die ECON-Befragung durchgeführt. ECON ist eine Studie zum ökonomischen Wissen und Einstellungen von Schüler*innen in NRW mit einem Wissenstest von 60 Minuten und einem Fragebogen von 25 Minuten. Als eigenständige Studie werden die ECON-Ergebnisse in einem eigenen Berichtsband aufgearbeitet.

In NRW fand zusätzlich eine Befragung zum ökonomischen Wissen und Einstellungen statt.

Lehrkräfte und Schulleitungen nahmen über Online-Fragebögen teil.

Die Befragung der Lehrkräfte und Schulleitungen erfolgte ausschließlich online. Die Anschreiben zum Lehrkraftfragebogen (TcQ) enthielten *Login*-Daten für eine *online* Plattform, auf der Lehrkräfte anonym Fragen zu ihrer Schule und auf politische Bildung und Demokratieerziehung bezogene Aktivitäten, zu darauf bezogenen Einstellungen sowie zu ihrer Ausbildung beantworten sollten. Der Lehrkräftefragebogen richtete sich in NRW und SH an zwei Zielgruppen: Die internationale Stichprobe und die nationale Ergänzungsstichprobe. Die internationale Stichprobe umfasst alle regulären Lehrkräfte, die in der 8. Jahrgangsstufe einer teilnehmenden Schule unterrichten, unabhängig davon, ob die unterrichtete Klasse für die Befragung ausgewählt wurde oder nicht. Die nationale Ergänzungsstichprobe umfasst zusätzlich alle Lehrkräfte, die an der Schule keine achte Klasse, aber ein Fach mit Bezug zu politischer Bildung und Demokratieerziehung unterrichten. In NRW zählten hierzu Lehrkräfte der Fächer Gesellschaftslehre, Wirtschaft und Arbeitswelt, Politik, Wirtschaft oder Wirtschaft-Politik. In SH waren es Lehrkräfte, die die Fächer Wirtschaft-Politik, Weltkunde, Verbraucherbildung und/oder Wirtschaftslehre unterrichteten. Zusätzlich wurden Schulleitungen gebeten, online einen kompakten Fragebogen (ScQ) zu Kontextinformationen der Schule auszufüllen, also beispielsweise die Größe der Schule, deren Urbanität und kulturelle Angebote im Umfeld. Zudem wurden auf politische Bildung und Demokratieerziehung bezogene Aktivitäten an der Schule erfragt.

Das nationale Studienzentrum selbst hat einen *National Context Questionnaire* (NCQ) ausgefüllt. Dieser ermöglicht einen ausführlichen Überblick über die

Struktur der verschiedenen teilnehmenden Bildungssysteme und deren Ansätze in der politischen Bildung und Demokratieerziehung.

Qualtitätssicherung

Um die Qualität der Erhebungen vor Ort an den Schulen sicherzustellen, haben sowohl Mitglieder des nationalen Forschungszentrums als auch unabhängige Beobachter*innen ausgewählte Erhebungen beobachtet und protokolliert. Die Beobachter*innen des nationalen Studienzentrums werden als *National Quality Observer* (NQO) bezeichnet und die externen Beobachter*innen als *International Quality Observer* (IQO). Die IQOs hatten auch die Aufgabe, die finalen Übersetzungen zu lesen und zu diesen Stellung zu beziehen.

Einige Erhebungen wurden beobachtet, um die Qualität sicherzustellen

Beobachtungen der Erhebungen wurden der Schulkoordination angekündigt. Die Testleiter*innen wurden nicht vorab informiert. Die Beobachtungen fanden sowohl in NRW als auch in SH statt. Es wurden systematisch unterschiedliche Schulformen besucht (Gymnasien, Realschulen, Hauptschulen, Förderschulen und Gesamtschulen) und sowohl ländliche wie städtische Regionen abgedeckt. Insgesamt wurden 28 NQO-Beobachtungen durchgeführt. Die Beobachter*innen füllten einen standardisierten Beobachtungsbogen aus, in dem sie unterschiedliche Aspekte der Testung dokumentierten. Hierzu gehörte der sichere Transport der Testmaterialien, das korrekte Ausfüllen der Teilnehmerlisten, die genaue Einhaltung der vorgegebenen Zeiten für die Durchführung jedes Testteiles und die präzise Erklärung über die Durchführung der Studie an die Schüler*innen seitens der Testleiter*innen basierend auf dem Manual für die Erhebungsleiter*innen. Darüber hinaus notierten die Beobachter*innen, ob es Störungen während der Durchführung gab (z. B. Lärm, unruhige Schüler*innen), ob Schüler*innen zu spät angekommen waren oder die Erhebung früher verlassen mussten. Es wurde darauf geachtet, wie Testleiter*innen auf Fragen der Schüler*innen reagierten. Während des Wissenstests waren beispielsweise Begriffserklärungen und andere Formen von Hilfe nicht gestattet, während der anderen Fragebögen waren Begriffserläuterungen jedoch erlaubt. Nach den Erhebungen wurden kurze Interviews mit den Testleiter*innen und den Schulkoordinator*innen durchgeführt. Sie wurden u. a. gefragt, wie sie die Vorbereitung und Durchführung der Erhebung wahrgenommen hatten und ob sie Verbesserungsvorschläge an die IEA hätten. Alle Informationen wurden an die IEA zurückgemeldet. Die Beobachtungen dienen der Einschätzung und Sicherung der Qualität der erhobenen Daten sowie der Evaluation der Prozesse, um eine gute Zusammenarbeit mit Testleiter*innen und reibungslosen Erhebungsdurchführungen sicherzustellen oder zu verbessern.

Die Beobachter*innen berichteten von keinen schwerwiegenden Problemen.

Die IQO folgte dem gleichen Prozess, jedoch waren die Beobachter*innen nicht mit dem nationalen Studienzentrum assoziiert. Sowohl die NQO als auch die IQO kamen zu ähnlichen Ergebnissen: Die Erhebungen wurden in der Regel vollständig den vorgeschriebenen Prozeduren folgend durchgeführt und die Erhebungsleiter*innen reagierten kompetent auf unvorhergesehene Herausforderungen. Selten wurde von Störungen der Erhebungen von außen berichtet und einige Schüler*innen nutzten ihr Handy nach dem Abschluss eines Testteils. Vereinzelt berichteten Beobachter*innen einen unzureichenden diskriminierungssensiblen Sprachgebrauch der Testleiter*innen in Interaktionen außerhalb des Skriptes.

Alle in diesem Kapitel beschriebenen Prozesse dienen der Sicherung von Qualität und möglichst hoher Invarianz, also der Sicherstellung der internationalen Vergleichbarkeit der Ergebnisse. Die Prozeduren sichern ab, dass sich gefundene Unterschiede zwischen den Ländern auf tatsächliche Unterschiede in der

Die beschriebenen Methoden dienen der internationalen Vergleichbarkeit der Ergebnisse.

Konstruktausprägung in der Population zurückführen lassen und nicht auf verschiedene Messmethoden oder Erfassungssituationen. Die Gefahr für Kontexteffekte auf die Daten ist für sprachbasierte Instrumente größer als für mathematisch orientierte Tests (Kankaraš & Moors, 2014; Laschke & Blömeke, 2016), daher sind die Anforderungen an die Methoden von ICCS besonders hoch.

Dieser nationale Bericht legt einen Fokus auf den Vergleich mit europäischen Bildungssystemen.

Der vorliegende Bericht fokussiert den Vergleich der Ergebnisse in NRW und SH mit den europäischen Teilnahmeländern (vgl. Kap. 20). Ergebnisse der drei nicht europäischen Länder werden in der Regel ebenfalls berichtet, diese Länder gehen jedoch nicht in die Berechnung des europäischen Vergleichswertes mit ein und werden auch nicht mit ihm verglichen.

Literatur

El Masri, Y. H., Baird, J.-A. & Graesser, A. (2016). Language effects in international testing: the case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, *23*(4), 427–455. https://doi.org/10.1080/0969594X.2016.1218323

Hahn-Laudenberg, K. & Abs, H. J. (2017). Politisches Wissen und Argumentieren. In H. J. Abs & K. Hahn-Laudenberg (Hrsg.), *Das politische Mindset von 14-Jährigen: Ergebnisse der International Civic and Citizenship Education Study 2016* (1. Auflage, S. 77–111). Waxmann.

Johansson, S. (2016). International large-scale assessments: what uses, what consequences? *Educational Research*, *58*(2), 139–148. https://doi.org/10.1080/00131881.2016.1165559

Kankaraš, M. & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381–399. https://doi.org/10.1177/0022022113511297

Koršňáková, P., Dept, S. & Ebbs, D. (2020). Translation: The preparation of national language versions of assessment instruments. In H. Wagemaker (Hrsg.), *IEA Research for Education. Reliability and Validity of International Large-Scale Assessment* (Bd. 10, S. 85–111). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_6

Laschke, C. & Blömeke, S. (2016). Measurement of job motivation in TEDS-M: testing for invariance across countries and cultures. *Large-scale Assessments in Education*, *4*(1). https://doi.org/10.1186/s40536-016-0031-5

Niemann, D. (2015). PISA in Deutschland: Effekte auf Politikgestaltung und -organisationen. *Die Deutsche Schule*, *107*(2), 141–157.

Rutkowski, L. & Rutkowski, D. (2019). Methodological challenges to measuring heterogeneous populations internationally. In L. Suter, E. Smith & B. D. Denman (Hrsg.), *The SAGE Handbook of Comparative Studies in Education.* Sage Publications.

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

Stubbe, T., Krieg, M., Beese, C. & Jusufi, D. (2020). Soziale Disparitäten in den mathematischen und naturwissenschaftlichen Kompetenzen von Viertklässlerinnen und Viertklässlern. In K. Schwippert, D. Kasper, O. Köller, N. McElvany, C. Selter, M. Steffensky & H. Wendt (Hrsg.), *TIMSS 2019: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 263–290). Waxmann Verlag GmbH.

van de Vijver, F. J. R., Jude, N. & Kruger, S. (2019). Challenges in international large-scale educational surveys. In L. Suter, E. Smith & B. D. Denman (Hrsg.), *The SAGE Handbook of Comparative Studies in Education* (S. 83–102). Sage Publications.

Weis, M., Müller, K., Mang, J., Heine, J.-H., Mahler, N. & Reiss, K. (2019). Soziale Herkunft, Zuwanderungshintergrund und Lesekompetenz. In K. Reiss, M. Weis, E. Klieme & O. Köller (Hrsg.), *PISA 2018.* Waxmann Verlag GmbH.

Bildnachweis S. 364:
Fotogruppe Die Bilderkammer Max-Brauer-Schule, Am Meer. Deutscher Jugendfotopreis 2014/DHM