Kapitel II
Ziele, Anlage und Durchführung der *Internationalen Grundschul-Lese-Untersuchung* (IGLU 2016)

Anke Hußmann, Heike Wendt, Daniel Kasper, Wilfried Bos
und Martin Goy

Im Jahr 2001 wurde als Folgestudie der *International Association for the Evaluation of Educational Achievement* (IEA) *Reading Literacy Study* (RLS) aus den Jahren 1990/1991 die *Progress in International Reading Literacy Study* (PIRLS) ins Leben gerufen, die weiterhin in der Verantwortung der IEA steht. Die IEA ist ein unabhängiger, internationaler Zusammenschluss von Wissenschaftlerinnen und Wissenschaftlern, Forschungseinrichtungen und Regierungsstellen und führt seit 1959 international vergleichende Schulleistungsstudien durch. PIRLS gehört seit 2001 zu den Kernstudien der IEA, wird alle fünf Jahre durchgeführt und trägt in Deutschland den Namen *Internationale Grundschul-Lese-Untersuchung* (IGLU). Zentrales Anliegen der IEA ist es, mit dieser Studie langfristige Entwicklungen in den teilnehmenden Bildungssystemen unter Berücksichtigung curricularer Vorgaben und weiterer zentraler Rahmenbedingungen zu schulischen Lehr- und Lernumgebungen zu dokumentieren.

Deutschland nimmt seit 2001 regelmäßig auch in den Jahren 2006, 2011 und 2016 mit der vierten Jahrgangsstufe an der Studie teil (Bos et al., 2003; Bos et al., 2007; Bos, Tarelli, Bremerich-Vos & Schwippert, 2012). Neben der Berichterstattung aktueller Befunde aus IGLU 2016 in nationaler und internationaler Perspektive ermöglicht diese Beteiligung auch, den Blick auf Entwicklungen im Grundschulwesen seit 2001 zu richten. Die Teilnahme an IGLU erfolgt auf Beschluss der *Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland* (KMK) und in Vereinbarung mit der KMK und dem *Bundesministerium für Bildung und Forschung* (BMBF). Sie ist Teil der systematischen Beobachtung des deutschen Bildungssystems im Rahmen der Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring (KMK, 2015).

Als internationale Schulleistungsstudie stellt IGLU einen Vergleichsmaßstab für Schülerleistungen bereit, mit dem gezielt jene Vergleiche in den Blick ge-

nommen werden können, die aus deutscher Perspektive von Erkenntnisinteresse sind, beispielweise zu weiteren Mitgliedstaaten der Europäischen Union. Damit unterscheidet sich IGLU von anderen Studien, die ebenfalls als Teil der KMK-Gesamtstrategie durchgeführt werden, zum Beispiel der *IQB-Bildungstrend* in der Primarstufe, der zur Überprüfung der in den Bildungsstandards formulierten Kompetenzanforderungen Vergleiche zwischen den Ländern der Bundesrepublik Deutschland erlaubt (zuletzt: Stanat, Schipolowski, Rjosk, Weirich & Haag, 2017). IGLU stellt als standardisierte und wissenschaftlich verantwortete Studie eine wichtige Datenbasis für die erziehungswissenschaftliche Forschung bereit. Die gewonnenen Daten dienen als Grundlagen, bildungspolitisch relevante Fragestellungen untersuchen und damit Wissen über Faktoren erlangen zu können, die die Qualität und Entwicklung von Bildungssystemen beeinflussen.

1 Die Beteiligung Deutschlands an international vergleichenden Schulleistungsstudien zur Lesekompetenz am Ende der Grundschulzeit

IGLU ist als Trendstudie konzipiert. In IGLU 2016 wurden zum vierten Mal die Lesekompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe untersucht. Vorläuferstudie von IGLU ist die *Reading Literacy Study* (RLS), die zuvor – ebenfalls von der IEA – in den Jahren 1990/1991 durchgeführt wurde. Beteiligt hatten sich damals 31 Länder (Elley, 1992, 1994). Deutschland nahm mit repräsentativen Stichproben für die westlichen und für die östlichen Länder der Bundesrepublik teil (Lehmann, Peek, Pieper & van Stritzky, 1995). Das Leseverständnis wurde in dritten und achten Klassen getestet, anders als in IGLU stand zum damaligen Zeitpunkt noch nicht das Ende der Grundschulzeit im Fokus der Untersuchung. Erhoben wurden Kompetenzen beim Lesen erzählender, darlegender und dokumentierender Texte anhand verschiedener Lesetexte und überwiegend mit Aufgaben im *Multiple-Choice*-Format.

Im Jahr 2001 wurde zum ersten Mal PIRLS – beziehungsweise in Deutschland IGLU – als international vergleichende Schulleistungsstudie durchgeführt. Die IEA stützte sich beim Aufbau des Erhebungsdesigns im Wesentlichen auf die Erfahrungen und Erkenntnisse aus der RL-Studie, setzte jedoch auch deutliche Modifikationen um. Das Leseverstehen wurde anhand zweier Leseintentionen und vier Leseverstehensprozessen untersucht (siehe Kapitel 3 in diesem Band). Um einen direkten Vergleich zwischen Ergebnissen aus der RL-Studie von 1990/91 und PIRLS/IGLU zu erreichen, wurden in neun Teilnehmerstaaten im Rahmen der Erhebung von PIRLS 2001 wiederholt Lesetests der RL-Studie eingesetzt (Martin, Mullis, Gonzalez & Kennedy, 2003). Deutschland war an dieser vergleichenden Untersuchung nicht beteiligt, die Ergebnisse aus IGLU 2001 und IGLU 2006 konnten jedoch mit jenen der RL-Studie kontrastiert werden (Bos et al., 2007).

An IGLU 2001 beteiligten sich international 35 Staaten und Regionen. In Deutschland wurden die eingesetzten Lesetests in 12 der 16 Länder um weitere Inhaltsbereiche ergänzt: An einem zweiten Testtag wurden mathematische, naturwissenschaftliche, orthographische Kompetenzen und die Schreibkompetenz der Schülerinnen und Schüler erhoben. Die Aufgaben zur Testung mathematischer und naturwissenschaftlicher Kompetenzen wurden zum Teil aus dem Aufgabenpool der *Trends in International Mathematics and Science Study*

(TIMSS 1995) für die Primarstufe entnommen, so dass eine internationale Verankerung gelang. Mit dieser nationalen Erweiterung wurden verlässliche Befunde zu Kompetenzen gewonnen, die am Ende der Grundschulzeit von Bedeutung sind.

Darüber hinaus entschieden sich in IGLU 2001 einige der Länder der Bundesrepublik Deutschland für eine Erweiterung der Stichprobe (*oversampling*), um repräsentative Ergebnisse für einzelne Länder zu erhalten und damit Vergleiche zwischen diesen Ländern (Baden-Württemberg, Bayern, Brandenburg, Bremen und Hessen) durchführen zu können. Für Nordrhein-Westfalen waren Analysen auf Länderebene möglich, weil im Rahmen der regulären Stichprobenziehung für Deutschland aus diesem Land ausreichend Schulen gezogen wurden. In Thüringen ergänzte das Ministerium des Landes die Stichprobe um weitere Schulen, mit dem Ziel, Schulentwicklungsprozesse unterstützen zu können. Bei der internationalen Auswertung wurden alle zusätzlich einbezogenen Schulen nicht berücksichtigt, sie dienten ausschließlich nationalen Analysen, wie detailliert dargelegt ist (Bos et al., 2003; Bos et al., 2004; Bos, Lankes, Prenzel, Valtin & Walther, 2005).

Die zweite Erhebung von IGLU wurde im Jahr 2006 durchgeführt. Beteiligt waren 45 Staaten und Regionen. Die Stichprobe in Deutschland wurde erweitert, so dass erstmalig Vergleiche zwischen allen Ländern möglich wurden. Pro Land wurden 25 Schulen gezogen, mit Ausnahme von Nordrhein-Westfalen, wo die Daten von 35 Schulen einbezogen wurden. Wie in IGLU 2001 so wurden auch in IGLU 2006 die internationalen um nationale Testkomponenten ergänzt, in 2006 um solche zur Erfassung orthographischer Kompetenzen. Da für Deutschland bereits eine Teilnahme an TIMSS 2007 (Bos et al., 2008) festgelegt war, wurde in 2006 auf die zusätzliche Testung mathematischer und naturwissenschaftlicher Kompetenzen verzichtet. Ausführliche Berichte zu den Ergebnissen der internationalen und nationalen Analysen liegen vor (Bos et al., 2007; Mullis, Martin, Kennedy & Foy, 2007).

Die dritte Erhebung von IGLU fand im Jahr 2011 statt. In Deutschland wurde sie zeitgleich mit TIMSS 2011 durchgeführt. Beteiligt mit der vierten Jahrgangsstufe (oder der äquivalenten Klassenstufe) waren 45 Staaten und Regionen als reguläre Teilnehmer. Erstmalig wurden in IGLU 2011 auch neun zusätzliche Regionen als sogenannte Benchmark-Teilnehmer aufgenommen, die ihre Ergebnisse zwar auf der internationalen Leistungsskala verorten möchten, deren Leistungsdaten aber nicht in die Berechnung des internationalen Mittelwerts eingehen. Aufgrund der zeitgleichen Durchführung von IGLU und TIMSS, entschied man in Deutschland (sowie in 36 anderen Teilnehmerstaaten und -regionen) in einer gemeinsamen Stichprobe zu erheben. Die Umsetzung erfolgte an zwei Testtagen, an denen in der einen Hälfte der Schulen die IGLU-Testkomponente am ersten Tag und die TIMSS-Testkomponente am zweiten Tag durchgeführt wurde, während in der zweiten Hälfte der Schulen die Testungen in umgekehrter Reihenfolge stattfanden. Getestet wurden sowohl Lesekompetenzen (IGLU) als auch mathematische und naturwissenschaftliche Kompetenzen (TIMSS). Die Stichprobe umfasste rund 200 Schulen. Die 16 Länder der Bundesrepublik Deutschland wurden gemäß ihrer Größe und Schulanzahl adäquat durch die Stichprobe abgebildet. Die nationalen und internationalen Befunde sind umfassend dokumentiert (Bos, Tarelli et al., 2012; Bos, Wendt, Köller & Selter, 2012; Mullis, Martin, Foy & Drucker, 2012; Wendt, Stubbe, Schwippert & Bos, 2015).

Lesekompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe im Trend zu untersuchen, setzt eine einheitliche Linie von Zyklus zu Zyklus voraus. Jeder Erhebungszyklus zeichnet sich jedoch durch Besonderheiten beziehungsweise Veränderungen aus. Für IGLU 2016 sind im Vergleich zu den vorangegangenen Erhebungszyklen folgende Spezifika zu benennen:

- Für die Einordnung der Leistungsergebnisse von Grundschulkindern steht nicht nur der internationale Bezugsrahmen zur Verfügung, sondern auch der Vergleich im Trend der Erhebungen aus 2001, 2006 und 2011. Damit sind erstmalig auf solider Datengrundlage Trendaussagen zu Kompetenzveränderungen am Ende der Grundschulzeit in den letzten 15 Jahren möglich.
- Im Gegensatz zu IGLU 2001 und IGLU 2006 lässt die Datengrundlage von IGLU 2016 keine Analyse auf der Ebene der Länder der Bundesrepublik Deutschland zu. Dies ist bereits seit IGLU 2011 nicht mehr möglich. Für Analysen zu Lesekompetenzen am Ende der Grundschulzeit im Vergleich der Länder sei daher auf den aktuell vorliegenden *IQB-Bildungstrend* 2016 verwiesen (Stanat et al., 2017).

2 Zentrale Erkenntnisse und Fragestellungen

Die einleitend beschriebene Zielsetzung von IGLU, den Ertrag von Bildungssystemen in international vergleichender Perspektive zu dokumentieren, wird hinsichtlich ihrer Möglichkeiten und Grenzen vielfältig diskutiert und reflektiert (z. B. Baumert, 2016; BMBF, 2001; Bos, Postlethwaite & Gebauer, 2010; Klieme, 2013). Herausgestellt werden in dieser Auseinandersetzung beispielsweise die verschiedenen Funktionen und Ansprüche, die mit Schulleistungsstudien verbunden sind (z. B. Howie & Plomp, 2005). Nach Klieme und Vieluf (2013) sind Schulleistungsstudien hinsichtlich ihrer Funktionen zu unterscheiden, die sie für unterschiedliche gesellschaftliche Gruppen erfüllen, zum Beispiel für die Politik, die Öffentlichkeit oder die Forschung. Aus dieser Perspektive betrachtet liefern Schulleistungsstudien wie IGLU als Systemmonitoringstudien *Indikatoren*, anhand derer sich die Strukturen, Funktionen und Erträge ebenso wie Disparitäten in Leistungen und die Bildungsteilhabe in Bildungssystemen beobachten, beschreiben und in einen internationalen Referenzrahmen einordnen lassen. Die Repräsentativität und die damit verbundene Aussagekraft der *Daten* erlauben zugleich, erziehungswissenschaftliche Grundlagenforschung zu betreiben, ebenso wie bildungspolitisch relevante Fragestellungen zu untersuchen. Es wird *Wissen* über Faktoren generiert, die die Qualität und Entwicklung von Bildungssystemen – sowohl in international vergleichender als auch in nationaler Perspektive – beeinflussen.

Mit der Durchführung international vergleichender Schulleistungsstudien wie IGLU verfolgt die IEA somit nicht nur die Feststellung dessen, was Bildungssysteme leisten. Zugleich werden potentielle Einflussfaktoren des Kompetenzerwerbs berücksichtigt, um weitere Erkenntnisse über die Bedingungen und Möglichkeiten der Verbesserung schulischer Förderung zu erhalten (IEA, 2012). Als potentielle Einflussfaktoren werden zentrale Rahmenbedingungen des Lesekompetenzerwerbs und des Unterrichts in der Grundschule untersucht (siehe Abbildung 2.1, Abschnitt 4.1.2). Die Lesekompetenzen der Schülerinnen und Schüler werden dazu anhand von Leistungstests ermittelt. Die Erfassung der Rahmenbedingungen erfolgt auf der Grundlage einer schriftlichen Befragung der Schülerinnen und Schüler, ihrer Eltern, der Lehrkräfte und der Schulleitungen so-

wie anhand der Einschätzung von Experten (siehe Abschnitt 5.2). Insofern bietet IGLU vielfältige Erkenntnismöglichkeiten, sowohl für Staaten, die regelmäßig an der Studie teilnehmen (z. B. um Leistungsveränderungen im Zeitverlauf zu beobachten), als auch für Staaten, die zum ersten Mal an IGLU teilnehmen (z. B. für den Leistungsvergleich mit anderen Staaten). IGLU gibt einen Einblick in die Bedingungen des Lehrens und Lernens am Ende der Grundschulzeit und trägt so zum Verständnis innerhalb und zwischen Bildungssystemen und ihrer Erträge bei. Aus diesen Erkenntnissen lassen sich Hinweise für Verbesserungen sowie weitergehende Forschungsbedarfe ableiten.

Die Teilnahme an IGLU 2016 und die damit verbundene Erhebung von umfassenden empirischen Daten erlauben repräsentative Rückschlüsse auf die Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe mit folgendem Erkenntnisgewinn:

- Die standardisierte Durchführung und die statistisch angemessene Auswertung elaborierter und erprobter Kompetenztestungen für den Bereich Leseverständnis lassen eine zuverlässige Einschätzung des Leistungsniveaus von Viertklässlerinnen und Viertklässlern in Deutschland im internationalen Vergleich zu.
- Die vierte Teilnahme an IGLU ermöglicht, Trends zu den Lesekompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe im internationalen Vergleich einzuschätzen. Die Auswertung erlaubt für die vergangenen 15 Jahre eine Bilanz der Entwicklungen der Leseleistungen am Ende der Grundschulzeit und der Rahmenbedingungen, in die der Lesekompetenzerwerb eingebettet ist.
- Die Befragung von Schülerinnen und Schüler zu ihrem Leseselbstkonzept, ihrer Lesemotivation und zu ihrem Lernverhalten ermöglicht eine zuverlässige Einschätzung der Bedeutsamkeit dieser Aspekte für die Leistungen von Viertklässlerinnen und Viertklässlern in Deutschland im internationalen Vergleich.
- Die Befragung von weiteren schulischen Akteurinnen und Akteuren (z. B. Fachlehrkräften und Schulleitungen) zu zentralen Bedingungsfaktoren schulischen Lernens ermöglicht eine zuverlässige Einschätzung der Bedeutsamkeit dieser Aspekte für die Testleistungen von Viertklässlerinnen und Viertklässlern in Deutschland im internationalen Vergleich.
- Die Befragung von Eltern, deren Kinder an IGLU teilgenommen haben, zu individuellen Voraussetzungen ermöglicht eine zuverlässige Einschätzung der Bedeutsamkeit dieser Aspekte für die Lesekompetenzentwicklung von Viertklässlerinnen und Viertklässlern in Deutschland im internationalen Vergleich.
- Die Stichprobenziehung ermöglicht, belastbare Aussagen zum Leistungsniveau und zur Bedeutsamkeit von individuellen Lernvoraussetzungen und Kontextfaktoren für Subpopulationen zu treffen (z. B. hinsichtlich der Aspekte Geschlecht, soziale Herkunft und Migrationshintergrund).

Der vorliegende Berichtsband stellt die Ergebnisse des internationalen Vergleichs aus deutscher Perspektive zu IGLU 2016 dar. Konzeptionelle und technische Besonderheiten sind für die Zielgruppe der Schülerinnen und Schüler am Ende der vierten Jahrgangsstufe dokumentiert. Vertiefende Fragestellungen werden in Nachfolgepublikationen aufgegriffen und behandelt. Im Fokus dieses Berichts steht die Beantwortung folgender Fragestellungen:

1. Welches Lesekompetenzniveau erreichen Schülerinnen und Schüler am Ende der vierten Jahrgangsstufe in Deutschland im Jahr 2016? Wie lassen sich die Ergebnisse im internationalen Vergleich bewerten? Haben sich die Ergebnisse seit 2001 verändert?
2. Wie lassen sich die IGLU-Leistungskennwerte auf Kompetenzstufen einordnen? Wie groß sind die Gruppen der auffällig leistungsschwachen und leistungsstarken Schülerinnen und Schüler? Wie unterscheiden sich die Ergebnisse von IGLU 2001, 2006, 2011 und 2016 diesbezüglich?
3. Welche Ergebnisse erzielen die Viertklässlerinnen und Viertklässler in Deutschland differenziert nach Textsorten und Verstehensprozessen? Wie unterscheiden sich die Ergebnisse von IGLU 2001, 2006, 2011 und 2016? Sind beobachtbare Veränderungen für bestimmte Subgruppen konstant?
4. Welche Bedeutung haben die individuellen Lernvoraussetzungen und Kontextfaktoren für die Lesekompetenzen? Welche Veränderungen von Lehr- und Lernbedingungen lassen sich seit 2001 beobachten?

3 IGLU 2016 – ein kooperatives Forschungsprojekt

Die Realisierung internationaler Vergleichsstudien wie IGLU ist aufwendig und erfordert die Zusammenarbeit verschiedener Organisationen, Institutionen und Personen auf nationaler und internationaler Ebene. Die IEA hat das *TIMSS & PIRLS International Study Center* (ISC) am *Boston College* in Chestnut Hill, Massachusetts, USA mit dem internationalen Management der Studie beauftragt. Es steht unter der Leitung von Ina V. S. Mullis, Professor an der *Lynch School of Education*, *Boston College*, sowie Michael O. Martin, Research Professor an der *Lynch School of Education, Boston College*. Das ISC verantwortet das Design und die Implementation der Studie, die internationale Koordination der Entwicklung der Instrumente und der Erhebungsprozeduren sowie die Qualität der Datenerhebung. Die internationale Skalierung (siehe Abschnitt 9) wird am ISC durchgeführt, auch der internationale Ergebnisbericht wird dort verfasst. Für Stichprobenziehung, Dokumentation der nationalen Stichproben und Berechnung der internationalen Stichprobengewichte kooperiert die internationale Studienleitung am ISC mit *Statistics Canada* in Ottawa, Ontario und der Abteilung der Stichprobenziehung an der *IEA Hamburg*.

Verantwortung für die Vorbereitung und Durchführung der Studie in jedem der an IGLU 2016 beteiligten Staaten trägt ein nationaler Projektkoordinator (der sogenannte *National Research Coordinator*, NRC). International vorgegebene Richtlinien sind dabei verbindlich. In Deutschland obliegt Prof. Dr. Wilfried Bos am *Institut für Schulentwicklungsforschung* (IFS) an der *Technischen Universität Dortmund* diese Verantwortung. Diese Aufgabe hat er gemeinsam mit Frau Dr. Heike Wendt, ebenfalls am IFS, wahrgenommen. Die Durchführung von IGLU in Deutschland wird zu gleichen Teilen vom BMBF und von der KMK finanziert.

Für die Analyse der Studienergebnisse und die Berichtslegung in Deutschland ist ein nationales Konsortium unter der Federführung des IFS verantwortlich, dem folgende Wissenschaftlerinnen und Wissenschaftler angehören:

Prof. Dr. Wilfried Bos	(NRC, Wissenschaftlicher Leiter von IGLU 2016 in Deutschland und Sprecher des Konsortiums) – Professur für Bildungsforschung und Qualitätssicherung am Institut für Schulentwicklungsforschung (IFS) an der Technischen Universität Dortmund
Dr. Heike Wendt	(NRC und Co-Projektleitung) – Akademische Rätin am Institut für Schulentwicklungsforschung (IFS) an der Technischen Universität Dortmund
Dr. Anke Hußmann	(Projektleitung) – Wissenschaftliche Mitarbeiterin am Institut für Schulentwicklungsforschung (IFS) an der Technischen Universität Dortmund
Prof. Dr. Albert Bremerich-Vos (i.R.)	(Mitglied des Konsortiums) – bis 2017 Professur für Linguistik und Sprachdidaktik an der Universität Duisburg-Essen
Prof. Dr. Nele McElvany	(Mitglied des Konsortiums) – Professur für Empirische Bildungsforschung mit dem Schwerpunkt Lehren und Lernen im schulischen Kontext an der Technischen Universität Dortmund, Geschäftsführende Direktorin des Instituts für Schulentwicklungsforschung (IFS)
Prof. Dr. Eva-Maria Lankes	(Kooptiertes Mitglied des Konsortiums) – Professur für Schulpädagogik an der Technischen Universität München, School of Education sowie Leitung der Qualitätsagentur des Staatsinstituts für Schulqualität und Bildungsforschung München
Prof. Dr. Tobias C. Stubbe	(Kooptiertes Mitglied des Konsortiums) – Professur für Schulpädagogik und Empirische Schulforschung am Institut für Erziehungswissenschaft an der Georg-August-Universität Göttingen
Prof. Dr. Renate Valtin (i.R.)	(Kooptiertes Mitglied des Konsortiums) – bis 2008 Professur für Grundschulpädagogik an der Humboldt-Universität zu Berlin, Präsidentin der Europäischen Lesegesellschaften

Ein großer Teil der Projektarbeit erfolgte hauptverantwortlich am IFS unter der Leitung von Dr. Anke Hußmann und Dr. Heike Wendt. Die methodische Betreuung der Studie am IFS oblag Dr. Daniel Kasper. Zu den beteiligten wissenschaftlichen Mitarbeiterinnen und Mitarbeitern zählen: Martin Goy, Maike Hoeft (bis Januar 2016), Dr. Svenja Rieser (bis Oktober 2017), Daniel Scott Smith (bis Juni 2016), Dr. Ruven Stahns und Anna Vaskova (bis Mai 2016). Das Projektteam wurde von vielen studentischen Hilfskräften unterstützt, darunter Cihan Günes, Ruth Engel (bis Juni 2017), Frederike Joosten, Donieta Jusufi, Adriana Kilisch, Katharina Roth, Lisa Schmitt (bis Oktober 2017), Maximilian Schulz, Felix Senger (bis August 2016), Charlotte Siepmann und Katja Weber. Mit der Organisation der nationalen Datenerhebung und -verarbeitung sowie der Aufgabenkodierung hat das IFS die *IEA Hamburg* beauftragt.

4 Anlage und Durchführung von IGLU 2016

Im vorliegenden Bericht werden zentrale Ergebnisse aus IGLU 2016 für Schülerinnen und Schüler am Ende der vierten Jahrgangsstufe vorgestellt. Im Vordergrund stehen die Ergebnisse des internationalen Vergleichs. Diese werden ergänzt um auf Deutschland fokussierte Analysen zu weiterführenden Fragestellungen. In diesem Abschnitt wird ein Überblick über das der Studie zugrundeliegende Rahmenmodell und die untersuchten Kompetenzbereiche gegeben. Vorgestellt werden darüber hinaus die Bildungssysteme, die an IGLU 2016 teilgenommen haben, sowie jene, die bei Vergleichen im Trend berichtet werden. Weitere Informationen, die dem besseren Verständnis beim Lesen des Berichtes dienen, sind ebenfalls dokumentiert, darunter die Teilnahmemodalitäten zum aktuellen Erhebungszyklus und zu den vergangenen Erhebungszyklen in den Jahren 2001, 2006 und 2011.

4.1 Die Rahmenkonzeption der Studie

Unter Berücksichtigung von Rahmenbedingungen, die für das Lernen in der Schule von Bedeutung sind, untersucht IGLU im internationalen Vergleich Leseleistungen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe. Von besonderem Interesse ist dabei die Frage, wie erfolgreich Schülerinnen und Schüler in den Bildungssystemen der teilnehmenden Staaten und Regionen am Ende der vierten Jahrgangsstufe grundlegende Kompetenzen erworben haben und ob sich Teilgruppen von Schülerinnen und Schülern identifizieren lassen, die sich in ihren Kompetenzen systematisch voneinander unterscheiden. Die Studie ist folglich so angelegt, dass gesicherte Aussagen über die Kompetenzen von Schülerinnen und Schülern im internationalen Vergleich getroffen werden können. Die theoretischen und konzeptionellen Grundlagen der Studie sind umfassend in einem Rahmenkonzept formuliert, das Expertinnen und Experten aus unterschiedlichen Teilnehmerstaaten erarbeitet haben (Mullis, Martin & Sainsbury, 2015). Es wurde erstmals zu IGLU 2001 vorgestellt und im Verlauf der folgenden Zyklen kontinuierlich angepasst (Campell, Kelly, Mullis, Martin & Sainsbury, 2001; Mullis, Kennedy, Martin & Sainsbury, 2006; Mullis, Martin, Kennedy, Trong & Sainsbury, 2009).

Das Verständnis von Lesekompetenz ist durch die aus dem angloamerikanischen Raum stammende *Literacy*-Konzeption definiert (siehe Abschnitt 4.1.1). Der nationalen Berichterstattung liegt darüber hinaus ein theoretisches Rahmenmodell zugrunde, das der Beschreibung des Zusammenhangs von Schülerleistungen und deren Bedingungen dient (siehe Abschnitt 4.1.2).

4.1.1 *Literacy* als Rahmenkonzept für die Testentwicklung

Theoretisch und konzeptionell liegt dem Verständnis von Lesekompetenz in IGLU die angloamerikanische *Literacy*-Konzeption zugrunde. Es handelt sich um ein pragmatisches Konzept von Grundbildung, in dem grundlegende Kompetenzen definiert sind, die in der Wissensgesellschaft von Bedeutung sind. Die Definition von *reading literacy*, auf die man sich in IGLU stützt, basiert auf einer Vielzahl an Theorien, in denen dieses Konzept als ein konstruktiver und

interaktiver Prozess verstanden wird (Anderson & Pearson, 1984; Chall, 1983; Kintch, 1998, 2012, 2013; Ruddell & Unrau, 2004; Rumelhart, 1985):

> „Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment." (Mullis et al., 2015, S. 12)

Die in IGLU erfasste Lesekompetenz wird im Sinne von *reading literacy* als die Fähigkeit verstanden, solche schriftsprachlichen Formen verstehen und nutzen zu können, die gesellschaftlich erforderlich und/oder für den Einzelnen von Bedeutung sind. Sie beschreibt demnach die Fähigkeit, in unterschiedlichen, für die Lebensbewältigung praktisch bedeutsamen Verwendungssituationen lesen zu können. Betont werden bei diesem Verständnis von Lesen die kulturelle Bedeutung von Bildungsinhalten, Partizipationsmöglichkeiten und die Entwicklung von Kompetenzen für ein selbstgesteuertes und lebenslanges Lernen. Es wird davon ausgegangen, dass junge Leserinnen und Leser auf unterschiedliche Weise die Bedeutung eines Textes konstruieren und das Leseverständnis im Zusammenhang mit Textsorten und Leseanlässen steht. Kinder lesen also, um zu lernen, um sich mit anderen Leserinnen und Lesern in der Schule und im Alltag auszutauschen und zu ihrem Vergnügen (ebd.). In IGLU werden drei Bereiche der Lesekompetenz betrachtet (siehe vertiefend die Ausführungen in den Kapiteln 3 und 4 in diesem Band):

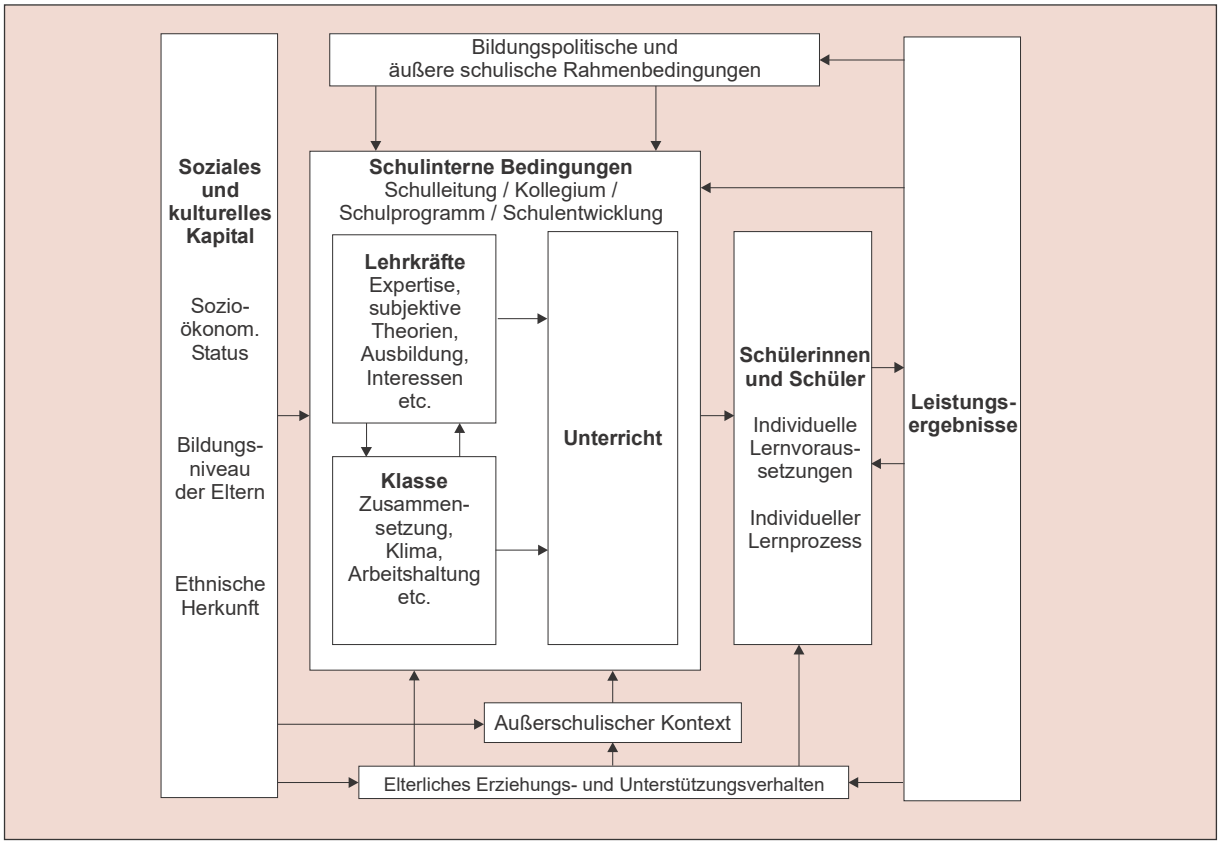
- Leseintentionen,
- Leseverstehensprozesse,
- Leseselbstkonzept, Lesemotivation und Leseverhalten.

Intentionen und Verstehensprozesse des Lesens von Schülerinnen und Schülern werden mit den eingesetzten Kompetenztests erhoben, Angaben zum Leseselbstkonzept, zur Lesemotivation und zum Leseverhalten und zu Einstellungen der Schülerinnen und Schüler mittels Fragebögen.

4.1.2 Rahmenmodell der Bedingungen schulischer Leistungen

Bereits zur nationalen Berichtslegung zu IGLU 2001 wurde ein theoretisches Rahmenmodell vorgestellt, das der Beschreibung des Zusammenhangs von Schülerleistungen und deren Bedingungen dienen sollte. In diesem Rahmenmodell werden Lehren und Lernen im Kontext gesellschaftlicher Ausgangsbedingungen verortet (vgl. Baumert & Weiß, 2002; Lankes et al., 2003; Prenzel & Doll, 2002; Wang, Haeterl & Walberg, 1993; Weinert & Helmke, 1997). Über die Erhebungszyklen hinweg hat es sich als Rahmenmodell zur Beschreibung schulischer Leistungen bewährt. Berücksichtigt werden darin die vielfältigen gesellschaftlichen, institutionellen und individuellen Bedingungen, in die das Lehren und Lernen in Grundschulen eingebettet ist. Bis heute hat das Modell nichts von seiner Bedeutung und Aktualität eingebüßt. Auch im vorliegenden Bericht dient es als Orientierungs- und Beschreibungsrahmen, um die Ebenen, die potentiell Einfluss auf die Qualität der Leistungen ausüben, in ihrem Beziehungsgeflecht betrachten zu können. Das Modell veranschaulicht, dass die erzielten Leistungen der Schülerinnen und Schüler als Ergebnis komplexer

Abbildung 2.1: Rahmenmodell zur Betrachtung des Zusammenhangs von Schülerleistungen und deren Bedingungen



IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

Lernprozesse in schulischen und außerschulischen Kontexten zu verstehen sind. Berücksichtigt sind gesellschaftliche und institutionelle Rahmenbedingungen sowie individuelle und familiäre Voraussetzungen (siehe Abbildung 2.1).

Dargestellt sind im Zentrum des Modells zum einen die Schülerinnen und Schüler, deren individuellen Lernvoraussetzungen und Lernprozesse in direktem Zusammenhang mit Leistungen beziehungsweise Leistungsergebnissen zu betrachten sind. Zum anderen sind die schulinternen Rahmenbedingungen, die Lehrkräfte, die Klassen und der Unterricht von zentraler Bedeutung. Als weitere im Modell aufgeführte Bedingungsfaktoren werden außerschulische und familiäre Merkmale, der soziokulturelle Hintergrund und das elterliche Erziehungs- und Unterstützungsverhalten einbezogen. Berücksichtigt werden darüber hinaus institutionelle Merkmale wie die bildungspolitischen und äußeren schulischen Rahmenbedingungen. Markiert durch die Pfeile, sind im Modell auch Hinweise auf Wechselwirkungen zwischen den Faktoren gegeben: Ergebnisse dieser Studie, so ist beispielsweise zu entnehmen, wirken auf bildungspolitische und äußere schulische Rahmenbedingungen sowie individuelle Lernvoraussetzungen und -prozesse.

Für die Analyse und die statistische Betrachtung von Unterschieden in den Schülerleistungen bieten die im Modell aufgezeigten Bedingungsfaktoren und deren komplexe wechselseitige Verflechtungen wertvolle Anhaltspunkte. Deshalb werden in IGLU die institutionellen und gesellschaftlichen Rahmenbedingungen, von denen angenommen wird, dass sie die Bereitstellung und Nutzung schulischer Lerngelegenheiten beeinflussen, mit einer standardisierten schriftlichen Befragung der teilnehmenden Schülerinnen und Schüler, ihrer Eltern, der

unterrichtenden Lehrkräfte und der Schulleitungen erfasst. Die so erhobenen Rahmendaten ermöglichen in der vertiefenden Analyse, Zusammenhänge zwischen den Leistungsergebnissen und den genannten Hintergrundmerkmalen zu beleuchten. Aussichtsreich ist dies insbesondere im Hinblick auf die Beschreibung und Erklärung von geschlechtsbezogenen, sozialen oder kulturellen Disparitäten in den Schülerleistungen (siehe auch Kapitel 5, 6 und 7 in diesem Band).

Die Rahmenbedingungen des Lehrens und Lernens werden in IGLU nicht nur mit Hilfe der administrierten Fragebögen an die Schulleitungen, die Lehrerinnen und Lehrer, die Schülerinnen und Schüler und Eltern erfasst. Zur Interpretation und Einordnung der Befunde dienen weitere Instrumente als Informationsquellen (Hooper, Mullis & Martin, 2015):

Ergänzt wird die internationale Berichterstattung der Leistungsergebnisse in IGLU durch die sogenannte *Enzyklopädie* (Mullis, Martin, Goh & Prendergast, 2017), in der die Schulsysteme der einzelnen Teilnehmerstaaten und -regionen beschrieben sind. Den Kapiteln der Enzyklopädie liegen einheitliche Kriterien zugrunde und sie werden von den nationalen Studienkoordinatoren verfasst. Schwerpunkte der Kapitel sind die Beschreibung der Organisation und der Strukturen des jeweiligen Bildungssystems und der fachspezifischen Curricula. Tabellarische Darstellungen zu demographischen Charakteristika der beteiligten Bildungssysteme, Inhalte und der Aufbau von Curricula sowie Spezifika der Lehrerinnen- und Lehrerbildung sind weitere Bestandteile der Kapitel. Erfasst werden die in der Enzyklopädie präsentierten Informationen von der internationalen Studienleitung mit Hilfe von standardisierten Fragebögen (*Curriculum Questionnaire*), die von ausgewählten Fachexpertinnen und Fachexperten der Teilnehmerstaaten und -regionen bearbeitet werden.

4.2 Teilnahme und Teilnahmemodalitäten an IGLU 2016

4.2.1 Teilnehmende 2016

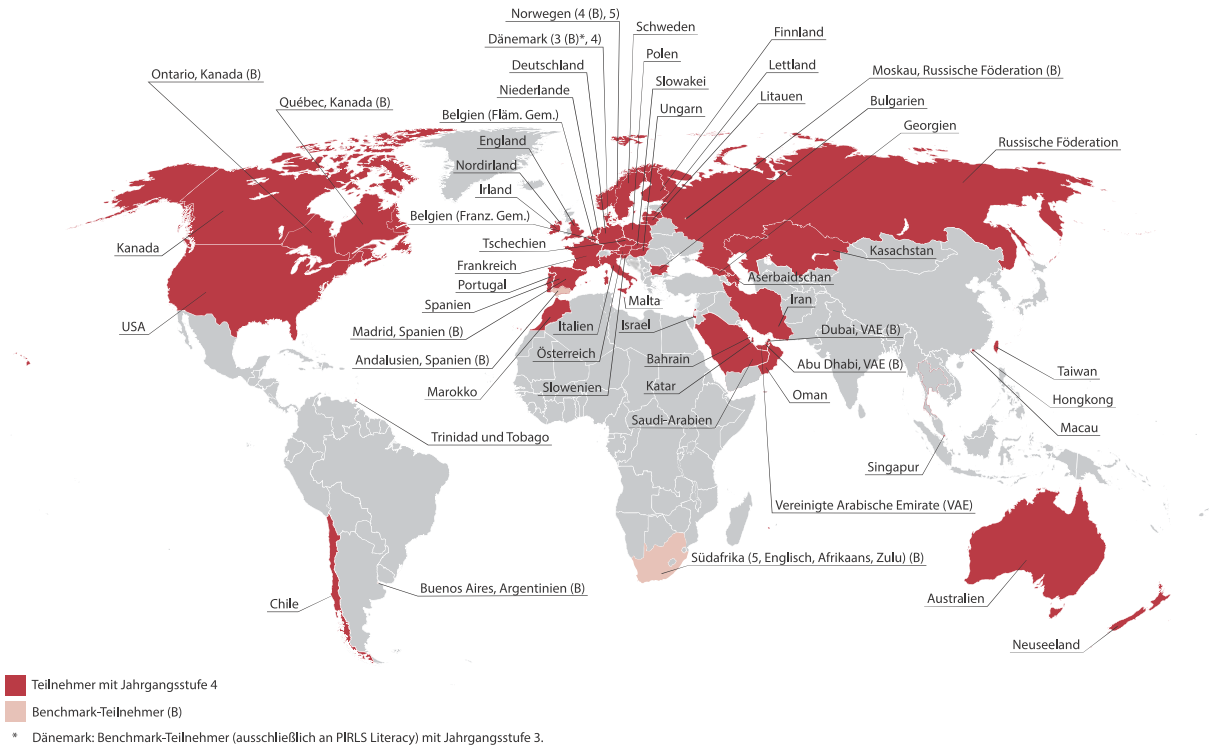
An IGLU 2016 haben sich weltweit 57 Bildungssysteme beteiligt. Es handelt sich überwiegend um souveräne Staaten. Teilgenommen haben auch Regionen, die als solche in Bildungsfragen weitgehend autonom handeln und deswegen in der internationalen Berichterstattung als Teilnehmer aufgeführt werden (z. B. die Flämische Gemeinschaft in Belgien). Insgesamt haben 47 Staaten und Regionen als reguläre Teilnehmer an der Studie teilgenommen (siehe Abbildung 2.2). Zehn sind Benchmark-Teilnehmer. Bei diesen handelt es sich um weitere Bildungssysteme, die ihre Ergebnisse auf der internationalen Ergebnisskala verorten möchten, deren Leistungsdaten aber nicht in die Berechnung des internationalen Mittelwerts eingehen.

Eine Besonderheit in IGLU 2016 stellt die Erweiterungsstudie PIRLS *Literacy* dar. In der Vergangenheit zeigte sich, dass in den Staaten, in denen die Lesekompetenzen der Viertklässlerinnen und Viertklässler deutlich unterhalb des internationalen Niveaus lagen, die Kompetenzen der leistungsschwächeren Kinder mit den regulären IGLU-Instrumenten nicht angemessen erfasst werden konnten. Für den Erhebungszyklus in 2016 entschied die IEA daher, IGLU um die Studienkomponente PIRLS *Literacy* zu erweitern (Mullis & Martin, 2015). PIRLS *Literacy* richtet sich auf Basiskompetenzen des Lesens, orientiert sich an der IGLU-Rahmenkonzeption und ist daher ähnlich wie IGLU aufgebaut. In PIRLS *Literacy* wurden zehn Lesetexte eingesetzt, von denen zwei (ein literari-

scher und ein informierender Text) auch in den regulären Testheften von IGLU rotieren (siehe Tabelle 2.3, Abschnitt 5.1.2). Drei Teilnehmer (Ägypten, Kuwait und Südafrika mit Jgst. 4) haben ausschließlich an PIRLS *Literacy* teilgenommen. Sie sind nicht Teil dieser Berichterstattung. Es gibt aber auch Teilnehmer, die als reguläre Teilnehmer an IGLU und zusätzlich an PIRLS *Literacy* teilgenommen haben (Iran und Marokko). Deutschland hat nicht an PIRLS *Literacy* teilgenommen, sondern lediglich als regulärer Teilnehmer an IGLU.

Kein Bestandteil dieses Berichts sind Ergebnisse der Teilnehmer aus PIRLS *Literacy*. Eine Ausnahme stellt Dänemark dar. Dänemark hat sich, neben der regulären Teilnahme mit der vierten Jahrgangsstufe, zusätzlich mit der dritten Jahrgangsstufe als Benchmark-Teilnehmer und hierbei ausschließlich an PIRLS *Literacy* beteiligt. Dieser Kohortenvergleich kann für Deutschland im europäischen Vergleich von Interesse sein. Daher wird Dänemark auf der Gesamtskala Lesen (siehe Kapitel 3, Abbildung 3.6 in diesem Band) als regulärer Teilnehmer mit der vierten Jahrgangsstufe und zusätzlich als Benchmark-Teilnehmer mit der dritten Jahrgangsstufe aufgeführt.

Abbildung 2.2: Staaten und Regionen, die an IGLU 2016 teilgenommen haben



In diesem Berichtsband sind die Ergebnisse der Schülerinnen und Schüler auf der Gesamtskala Lesen somit für alle beteiligten Staaten, Regionen und Benchmark-Teilnehmer, einschließlich Dänemark mit Jahrgangsstufe 3 als Benchmark-Teilnehmer an PIRLS *Literacy*, dokumentiert.

Wie in den vorangegangenen Erhebungszyklen werden bei der Darstellung der Ergebnisse Vergleichsgruppen (VG) gebildet. In diese Vergleichsgruppen werden jene Staaten und Staatengruppen aufgenommen, die für einen Vergleich mit Deutschland geeignet sind (z. B. aufgrund eines ähnlichen kulturellen Hintergrunds oder einer ähnlichen wirtschaftlichen Situation). Dabei handelt es sich

um die an IGLU 2016 teilnehmenden Mitglieder der *Europäischen Union* (EU) und der *Organisation for Economic Co-operation and Development* (OECD). In den Ergebnisdarstellungen sind in den nachfolgenden Kapiteln nicht nur Mittelwerte für einzelne Teilnehmer, sondern auch Mittelwerte für die beiden Vergleichsgruppen der EU (VG EU) und OECD (VG OECD) dargestellt. Die Teilnehmer der Vergleichsgruppen sind in Tabelle 2.1 aufgeführt.

In der Berichterstattung zum Erhebungszyklus von IGLU 2016 werden lediglich bei der Darstellung der Gesamtskala Lesen sämtliche der oben genannten Teilnehmer berücksichtigt. Da der Fokus auf den Leistungen von Schülerinnen und Schülern in Deutschland liegt, werden in den übrigen Ergebnisdarstellungen zu IGLU 2016 ausschließlich Ergebnisse derjenigen Staaten, Regionen und Benchmark-Teilnehmer dokumentiert, auf die eines der folgenden Kriterien zutrifft: (1) Teilnehmer ist Mitglied der EU, (2) Teilnehmer gehört der OECD an, (3) Teilnehmer, hat auf der Gesamtskala Lesen besser oder nicht signifikant von Deutschland verschieden abgeschnitten.

Tabelle 2.1: Vergleichsgruppen in IGLU 2016

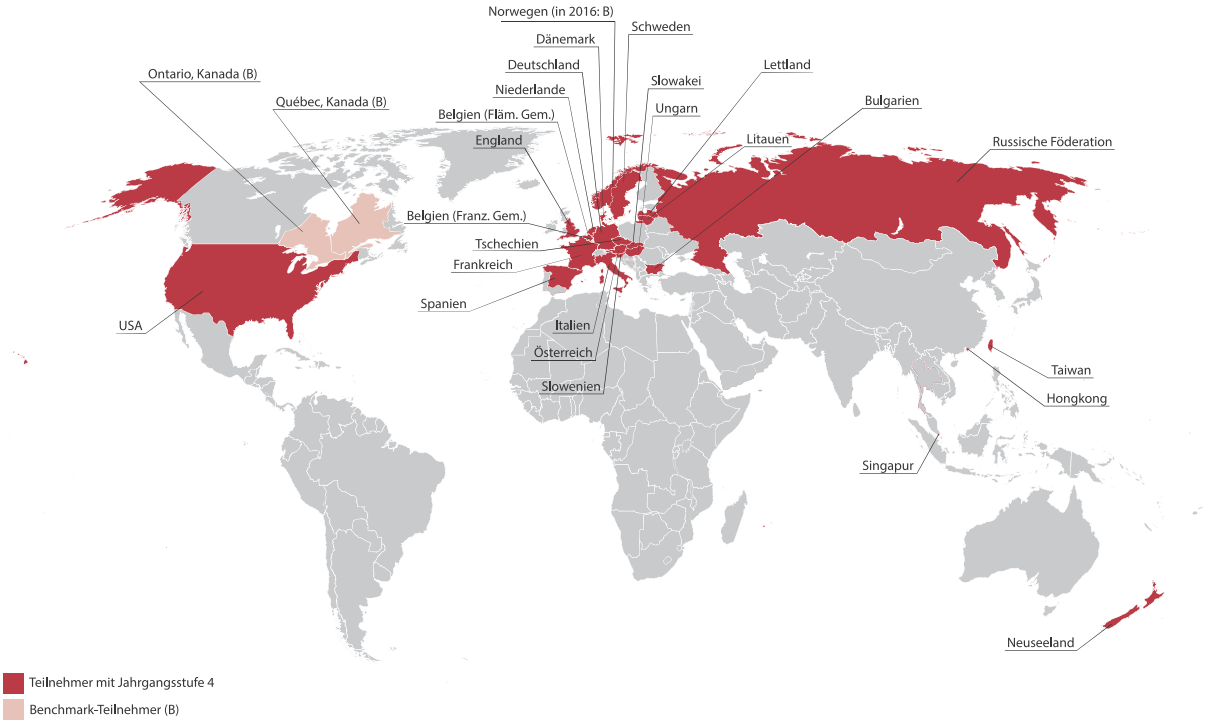
VG EU		VG OECD	
Belgien (Fläm. Gem.)	Nordirland	Australien	Neuseeland
Belgien (Franz. Gem.)	Österreich	Belgien (Fläm. Gem.)	Niederlande
Bulgarien	Polen	Belgien (Franz. Gem.)	Nordirland
Dänemark	Portugal	Chile	Norwegen
Deutschland	Schweden	Dänemark	Österreich
England	Slowakei	Deutschland	Polen
Finnland	Slowenien	England	Portugal
Frankreich	Spanien	Finnland	Schweden
Irland	Tschechien	Frankreich	Slowakei
Italien	Ungarn	Irland	Slowenien
Lettland		Israel	Spanien
Litauen		Italien	Tschechien
Malta		Kanada	Ungarn
Niederlande		Lettland	USA
	24		**28**

IEA: Progress in International Reading Literacy Study (PIRLS)

4.2.2 Teilnehmende IGLU 2001, 2006, 2011 und 2016

Einige Staaten, Regionen und Benchmark-Teilnehmer haben, wie Deutschland auch, an vergangenen Erhebungszyklen von IGLU teilgenommen. Für die Darstellung von Trends werden in diesem Bericht insgesamt 24 reguläre Teilnehmer und drei Benchmark-Teilnehmer betrachtet (siehe Abbildung 2.3).

Abbildung 2.3: Teilnehmer der Trendvergleiche zu IGLU 2001, 2006, 2011 und 2016



IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

Berücksichtigt werden diejenigen Bildungssysteme, (1) die zusätzlich zu IGLU 2016 an mindestens zwei weiteren IGLU-Zyklen teilgenommen haben und (2) zugleich Mitglied der EU und/oder der OECD sind und/oder auf der Gesamtskala Lesen in IGLU 2016 signifikant besser oder nicht signifikant verschieden von Deutschland abgeschnitten haben. Im Trendvergleich nicht berücksichtigt werden daher die Teilnehmer Georgien, Iran, Katar, Marokko, Trinidad und Tobago sowie Benchmark-Teilnehmer Südafrika (5. Jgst.). Ebenfalls im Trendvergleich nicht berücksichtigt werden Israel und Polen aufgrund von Änderungen in den Übersetzungen der Testinstrumente oder in den Erhebungsbedingungen. Eine Ausnahme ist bei der hier vorgenommenen Trendauswahl zu berücksichtigen: Die Flämische Gemeinschaft in Belgien wird in die Trendgruppe einbezogen, obwohl sie neben IGLU 2016 an lediglich einem weiteren Erhebungszyklus (IGLU 2006) teilgenommen hat. Grund hierfür ist der für den vorliegenden Berichtsband aus deutscher Perspektive relevante Vergleich.

4.2.3 Teilnahmebedingungen der Studie

Bildungssysteme miteinander systematisch zu vergleichen, setzt die Einhaltung strenger methodischer Standards voraus. In diesen Standards ist definiert, mit welchen Schülerinnen und Schülern und unter welchen Bedingungen sich einzelne Bildungssysteme an der Studie beteiligen dürfen (Martin, Mullis & Hooper, 2017). Die Teilnahmebedingungen, die international vorgegeben sind, werden im Folgenden für IGLU 2016 resümiert.

4.2.3.1 Definition der Untersuchungspopulation

Ziel von IGLU ist es, den Ertrag von Bildungssystemen in international vergleichender Perspektive festzustellen. Dazu werden systematisch Leistungen von Schülerinnen und Schülern der vierten Jahrgangsstufe in den Blick genommen. Angestrebt wird, komplette Schülerjahrgänge miteinander zu vergleichen. Für die Definition der zu untersuchenden Schülerpopulation (Zielpopulation) steht dabei – wie in den vorangegangenen Erhebungszyklen – zunächst das Kriterium der formalen Beschulungszeit im Vordergrund:

Für alle Teilnehmerstaaten sind als Zielpopulation alle Schülerinnen und Schüler definiert, die sich nach der *International Standard Classification of Education* im vierten Jahr formaler Beschulung befinden (ISCED Level 1, vgl. UIS, 2015). Dies trifft in den meisten Teilnehmerstaaten auf Schülerinnen und Schüler der vierten Jahrgangsstufe zu. Um ergänzend zu dieser Referenzkategorie eine entwicklungsgerechte Passung der Leistungstests und der Durchführungsbedingungen der Tests sicherzustellen, ist als zusätzliches Kriterium ein Durchschnittsalter von mindestens 9 Jahren und 6 Monaten der definierten Zielpopulation zum Testzeitpunkt bestimmt worden (Martin, Mullis & Foy, 2015).

Ausschöpfung der Zielpopulation

Internationale Vergleichbarkeit setzt unter anderem voraus, dass die definierte Zielpopulation in allen Teilnehmerstaaten möglichst vollständig ausgeschöpft wird. Als erfüllt gilt diese Vorgabe, wenn alle Schülerinnen und Schüler, die unter die Definition der internationalen Zielpopulation fallen, die Möglichkeit erhalten, an der Studie teilnehmen zu können. Ausgehend von der international definierten Zielpopulation muss jeder Teilnehmer die international vorgegebenen Richtlinien zur Stichprobenziehung an die Spezifika des eigenen Systems anpassen. Ein Handbuch zur Stichprobenziehung (Martin et al., 2017) dient dabei als grundlegender Orientierungsrahmen. Unterstützung bei der Umsetzung dieses Vorgehens bieten Expertinnen und Experten von *Statistics Canada* und der *IEA Hamburg*.

Manche nationalen Besonderheiten hindern einzelne Teilnehmer, den internationalen Vorgaben zu entsprechen. Beispielsweise gibt es in fast jedem Schulsystem einzelne Schulen oder Gruppen von Schülerinnen und Schülern, die nicht an der Erhebung teilnehmen können oder aus nationalen Gründen nicht teilnehmen sollen. Sogenannte *Ausschlüsse von der Zielpopulation* können nach eindeutig von der internationalen Studienleitung definierten Kriterien von jedem Teilnehmer vorgenommen werden:

Ausgeschlossen werden können auf *Schülerebene* solche Schülerinnen und Schüler, die aus körperlichen, emotionalen oder geistigen Gründen nicht in der Lage sind, den Test selbstständig zu bearbeiten, oder Schülerinnen und Schüler, die weniger als ein Jahr in der Testsprache unterrichtet wurden und deren Muttersprache nicht die Testsprache ist.

Auf *Schulebene* können jene Schulen ausgeschlossen werden, die in besonders schwer erreichbaren Regionen liegen, die weniger als vier Schülerinnen und Schüler im vierten Jahrgang beschulen (d. h. eine besonders geringe Schülerzahl aufweisen), die in Lehrplan oder Schulstruktur vom nationalen Schulsystem abweichen oder ausschließlich Schülerinnen und Schüler unterrichten, auf die die Schülerebene betreffenden Ausschlusskriterien zutreffen. Ausschlüsse dürfen auf

Schul- und auf Schülerebene insgesamt nicht mehr als 5 Prozent der international definierten Zielpopulation betragen (Joncas & Foy, 2011).

Unter Berücksichtigung dieser Ausschlusskriterien wird für jeden Teilnehmer die *nationale, effektiv erreichte Zielpopulation (national effective target population)* bestimmt. Diese beschreibt, welche Schülerinnen und Schüler einer Population tatsächlich durch die Stichprobe der teilnehmenden Schülerinnen und Schüler repräsentiert werden. Auftretende Abweichungen von der international definierten Zielpopulation in den internationalen Berichten sind auf der Grundlage dieser Beschreibung nachvollziehbar dokumentiert, bedeutsame Abweichungen sind hervorgehoben. Informationen zu den Ausschlüssen der Teilnehmer in IGLU 2016 sowie zu den Teilnehmern der vorangegangenen Erhebungszyklen in 2001, 2006 und 2011 finden sich unter anderem im Anhang A des vorliegenden Bandes.

4.2.3.2 Veränderte Teilnahmemodalitäten in Deutschland seit IGLU 2001

Veränderungen in der Zusammensetzung der Stichproben können unter anderem durch Ausschlüsse von der Zielpopulation begründet (siehe Abschnitt 4.2.3.1), oder auch Folge von Veränderungen der Teilnahmemodalitäten in den einzelnen Staaten und Regionen sein.

Im Vergleich der Erhebungszyklen von IGLU 2001, 2006, 2011 und 2016 zeigt sich, dass je nach Erhebungsjahr und je nach Land unterschiedliche Verpflichtungsgrade der Teilnahme bestehen. So war zum Erhebungszyklus im Jahr 2001 die Teilnahme an IGLU noch in allen Ländern der Bundesrepublik Deutschland freiwillig. Ein Kind konnte jederzeit die Bearbeitung der Tests ablehnen oder abbrechen, selbst wenn von den Eltern eine Einverständniserklärung vorlag (ein solcher Fall wurde jedoch nicht bekannt) (vgl. Lankes et al., 2003, S. 26).

Erst seit IGLU 2006 ist einheitlich definiert, dass die Teilnahme am Leistungstest für alle Schülerinnen und Schüler verpflichtend ist. Die Bearbeitung des Schülerfragebogens unterliegt jedoch bundesweit unterschiedlichen Regelungen (siehe Tabelle 2.2). Während die eine Hälfte der Länder die Befragung der Schülerinnen und Schüler (teil-)verpflichtend durchführt (Berlin, Brandenburg, Bremen, Hessen, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt und Thüringen), nehmen in der anderen Hälfte der Länder Schülerinnen und Schüler nach Vorlage der schriftlichen Einverständniserklärung ihrer Eltern freiwillig an der Befragung teil (Baden-Württemberg, Bayern, Hamburg, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland, Sachsen und Schleswig-Holstein).

Unterschiedliche Verpflichtungsgrade gelten auch in den einzelnen Ländern im Hinblick auf die Teilnahme an der schriftlichen Befragung der Lehrkräfte sowie der Schulleitungen. Die Rechtslage eines Landes kann Lehrkräfte und Schulleitungen zur Teilnahme an der Befragung in IGLU verpflichten, teilverpflichten oder zur freiwilligen Teilnahme lediglich ermuntern. Teilverpflichtungen schließen ein, dass Angaben zur Schule verpflichtend sind, Angaben zu personenbezogenen Daten hingegen freiwillig. Unverändert für alle Länder und über die Erhebungszyklen hinweg ist die freiwillige Teilnahme an der Elternbefragung.

Tabelle 2.2: Länderspezifische Modalitäten der Teilnahme an IGLU 2001, 2006, 2011 und 2016

Länder	Verpflichtungsgrade der Teilnahme an den Erhebungsinstrumenten an öffentlichen Schulen				
	Leistungstest[A]	Schüler-fragebogen[A]	Lehrkräfte-fragebogen[A]	Schul-fragebogen[A]	Eltern-fragebogen
Baden-Württemberg	verpflichtend[1]	freiwillig	freiwillig	freiwillig	freiwillig
Bayern	verpflichtend[1]	freiwillig	freiwillig	freiwillig	freiwillig
Berlin	verpflichtend	teilverpflichtend[2]	teilverpflichtend[2]	teilverpflichtend[2]	freiwillig
Brandenburg	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig
Bremen	verpflichtend	verpflichtend[3]	teilverpflichtend[4]	teilverpflichtend[4]	freiwillig
Hamburg	verpflichtend	freiwillig	freiwillig[5]	teilverpflichtend	freiwillig
Hessen	verpflichtend	verpflichtend	teilverpflichtend	teilverpflichtend	freiwillig
Mecklenburg-Vorpommern	verpflichtend	verpflichtend[6]	verpflichtend	verpflichtend	freiwillig
Niedersachsen	verpflichtend[1]	verpflichtend	verpflichtend	verpflichtend	freiwillig
Nordrhein-Westfalen	verpflichtend	freiwillig	verpflichtend	verpflichtend	freiwillig
Rheinland-Pfalz	verpflichtend	freiwillig	teilverpflichtend[7]	teilverpflichtend[7]	freiwillig
Saarland	verpflichtend	freiwillig	teilverpflichtend[8]	teilverpflichtend[8]	freiwillig
Sachsen	verpflichtend	freiwillig	freiwillig	freiwillig[9]	freiwillig
Sachsen-Anhalt	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig
Schleswig-Holstein	verpflichtend	freiwillig	freiwillig	freiwillig	freiwillig
Thüringen	verpflichtend	verpflichtend	verpflichtend	verpflichtend	freiwillig

A = Im Jahr 2001 war die Teilnahme an IGLU in allen Ländern der Bundesrepublik Deutschland freiwillig.
1 = In Baden-Württemberg, Bayern und Niedersachen war im Jahr 2006 die Teilnahme am Test freiwillig.
2 = In Berlin war in den Jahren 2006 und 2011 die Teilnahme am Schüler-, Lehrkräfte- und Schulfragebogen verpflichtend.
3 = In Bremen war in den Jahren 2006 und 2011 die Teilnahme am Schülerfragebogen freiwillig.
4 = In Bremen war im Jahr 2006 die Teilnahme am Lehrer- und Schulfragebogen verpflichtend.
5 = In Hamburg war in den Jahren 2006 und 2011 die Teilnahme am Lehrkräftefragebogen teilverpflichtend.
6 = In Mecklenburg-Vorpommern war im Jahr 2006 die Teilnahme am Schülerfragebogen freiwillig.
7 = In Rheinland-Pfalz war im Jahr 2006 die Teilnahme am Lehrer- und Schulfragebogen verpflichtend.
8 = Im Saarland war in den Jahren 2006 und 2011 die Teilnahme am Lehrkräfte- und Schulfragebogen verpflichtend.
9 = In Sachsen war im Jahr 2006 die Teilnahme am Schulfragebogen teilverpflichtend.

IEA: Progress in International Reading Literacy Study (PIRLS)

Für Schulen in privater Trägerschaft gilt mehrheitlich, dass die Teilnahme an der gesamten Studie freiwillig ist. In manchen Ländern wie Baden-Württemberg, Bayern oder Hamburg ist für diese Schulen lediglich die Teilnahme an den Leistungstests verpflichtend.

In Tabelle 2.2 sind die Verpflichtungsgrade der länderspezifischen Teilnahme an öffentlichen Schulen für die Erhebungszyklen von IGLU in den Jahren 2001, 2006, 2011 und 2016 zusammenfassend dargestellt. Die Tabelle dokumentiert die zum Zeitpunkt für 2016 geltenden Teilnahmemodalitäten pro Land und pro Erhebungsinstrument. Abweichungen in den Teilnahmebedingungen im Vergleich der Erhebungszyklen 2001, 2006 und 2011 sind durch Fußnoten mit entsprechenden Erläuterungen ausgewiesen. Insbesondere für Trendanalysen sind diese Informationen relevant, weil sie Hintergründe für mögliche veränderte Stichprobenzusammensetzungen und Teilnahmequoten der verschiedenen befragten Akteursgruppen liefern.

4.3 Verfahren und Kriterien der Stichprobenziehung

In IGLU erfolgt die Stichprobenziehung nach einem zweistufigen stratifizierten Clusterdesign (Martin et al., 2017; Joncas & Foy, 2011). Es wird eine zufällige Auswahl aus allen Schulen der national effektiv erreichten Zielpopulation gezogen und anschließend eine oder mehrere Klassen in den ausgewählten Schulen ermittelt. Die Auswahl der Schulen erfolgt in zwei Stufen:

In der *ersten Stufe* wird aus einer Liste (*Sampling Frame*), auf der alle Schulen verzeichnet sind, die Teil der national effektiv erreichten Zielpopulation sind, eine Zufallsstichprobe von Schulen gezogen. Die Mindestanzahl beträgt nach den Vorgaben der internationalen Studienleitung 150 Schulen. Da Schulen mit einer größeren Anzahl an Schülerinnen und Schülern in der vierten Jahrgangsstufe auch einen höheren Anteil an der Population haben, wird die Information zur Größe der Jahrgangsstufe 4 bei der Stichprobenziehung berücksichtigt. Zur Steigerung der Effizienz der Stichprobe und der Präzision der Ergebnisse wird die Methode der Stratifizierung angewendet: Schulen, die als Teil der national effektiv erreichten Zielpopulation gelten, werden nach bestimmten Merkmalen (*Strata*) kategorisiert (*geschichtet*). Mögliche *Strata* sind die geographische Lage (z. B. Länder der Bundesrepublik Deutschland) oder der Schultyp (z. B. Förder- vs. Regelschule, privat vs. öffentlich). Diese zur Verfügung stehenden Informationen über die Schulen werden bei der Stichprobenziehung berücksichtigt, um sicherzustellen, dass sich die tatsächliche Verteilung der Schulen innerhalb der *Strata* auch in der Stichprobe widerspiegelt und spezifische Gruppen der Population adäquat repräsentiert sind.

Innerhalb der in der ersten Stufe ausgewählten Schulen erfolgt in der *zweiten Stufe* die Ziehung der Klassenstichproben. Jede Klasse hat dieselbe Wahrscheinlichkeit, in die Stichprobe zu gelangen. Mit der Auswahl der Klassen sind automatisch auch die Schülerinnen und Schüler bestimmt, da – abgesehen von den oben beschriebenen Ausnahmen – alle Schülerinnen und Schüler einer Klasse am Test teilnehmen sollen. Nach internationalen Vorgaben sind mindestens 4000 Schülerinnen und Schüler auszuwählen (ebd.).

4.3.1 Schul- und Schülerteilnahmequoten

In jeder Studie ist mit Stichprobenausfällen zu rechnen. Um Datenverzerrungen aufgrund nicht teilnehmender Schulen oder Personengruppen in IGLU gering zu halten, sind seitens der internationalen Studienleitung Vorgaben für minimale Beteiligungsquoten definiert (Martin et al., 2017; Joncas & Foy, 2011). Teilnehmende Staaten und Regionen, die die vorgeschriebenen Rücklaufquoten nicht erreichen, werden in der internationalen Berichterstattung bei weniger gravierenden Abweichungen mit einer Fußnote und entsprechender Erklärung in den Ländervergleich einbezogen. Erweisen sich die Rücklaufquoten als zu gering, werden die Ergebnisse gesondert berichtet.

Auf *Schulebene* liegt die minimal zu erreichende Beteiligungsquote bei 85 Prozent der ursprünglich ausgewählten Schulen. Fehlen weniger als 15 Prozent der ursprünglich ausgewählten Schulen, können Schulen, die nicht teilgenommen haben, durch Ersatzschulen im Nachrückverfahren angesprochen werden. Dabei handelt es sich um Schulen, die bereits im Rahmen der Stichprobenziehung ausgewählt wurden. Dieses Vorgehen ermöglicht, die Anforderungen an die Stichprobengröße insgesamt erfüllen zu können. Ersatzschulen müssen den ur-

sprünglich gezogenen Schulen in den für die Stratifizierung bestimmten Merkmalen ähneln.

Auf *Klassenebene* liegt die minimal zu erreichende Beteiligungsquote bei 95 Prozent der ursprünglich ausgewählten Klassen. Für nicht teilnehmende Klassen müssen Klassen in Ersatzschulen gefunden werden. Es besteht in diesen Fällen nicht die Möglichkeit, diese durch Parallelklassen zu ersetzen. Auch für die Befragung der Lehrerinnen und Lehrer ist für die spätere Berichtslegung eine Mindestbeteiligung von 85 Prozent vorgegeben. Wie im Falle der Beteiligungsquote von Klassen, so dürfen auch Lehrkräfte den internationalen Vorgaben zufolge nicht ersetzt werden (ebd.).

Für die *Schülerebene* ist eine Beteiligungsquote von mindestens 85 Prozent der Schülerinnen und Schüler festgelegt. Zusätzlich gilt, dass die Beteiligung in den Klassen nicht unter 50 Prozent sinken darf. Um Rücklaufquoten im geforderten Maße zu erreichen, wird an Schulen ein Nachtest durchgeführt, an denen am Testtag mehr als 10 Prozent der Schülerinnen und Schüler fehlen. Wird trotz Nachtests die Mindestrücklaufquote nicht erreicht, wird die Klasse und damit auch die ganze Schule von der Studie ausgeschlossen.

Alternativ zu den genannten Beteiligungsquoten ist zu gewährleisten, dass die kombinierte Schul- und Schülerteilnahmequote bei mindestens 75 Prozent liegt, das heißt, der durch Ausschluss auf Schul-, Klassen- und Schülerebene bedingte Anteil an nicht teilnehmenden Schülerinnen und Schülern darf 25 Prozent nicht überschreiten.

Die Bewertung der Qualität der Stichprobe auf Schülerebene erfolgt auch danach, ob die Teilnahme an der Leistungsmessung zu verwertbaren Ergebnissen geführt hat. Als Interpretationsgrundlage dienen dabei die Kompetenzmittelwerte, anhand derer aufzuzeigen ist, dass für mindestens 85 Prozent der Schülerinnen und Schüler innerhalb der Stichprobe zuverlässige Leistungswerte geschätzt werden können. Die Schätzung des Leistungswertes gilt als zuverlässig, wenn eine Schülerin oder ein Schüler mehr Aufgaben richtig bearbeitet hat als durch Raten richtig zu lösen wären.

4.3.2 Besonderheiten der Stichproben im internationalen Vergleich

Die Tabelle A.4 in Anhang A dieses Bandes gibt einen Überblick über zentrale Kennwerte der Stichproben der an IGLU 2016 teilnehmenden Bildungssysteme. Zusätzlich sind hier auch die Angaben für die Teilnehmer an PIRLS *Literacy* aufgenommen. Die Tabelle dokumentiert Teilnahmen an den Erhebungszyklen von IGLU sowie Besonderheiten bezüglich der nationalen Zielpopulationen, der Schul- und Schülerteilnahmequoten sowie der Leistungsmessungen. Norwegen als Teilnehmer mit Jahrgangsstufe 5, die Benchmark-Teilnehmer sowie die Teilnehmer an PIRLS *Literacy* werden gesondert aufgeführt (siehe Abschnitt 4.2.1). Im Anhang A sind zudem in Tabelle A.1 diese Besonderheiten entsprechend für IGLU 2001, in Tabelle A.2 für IGLU 2006 und in Tabelle A.3 für IGLU 2011 dargestellt. In der ersten Spalte dieser Tabellen wird durch Ziffernfußnoten auf Besonderheiten der Stichproben verwiesen, zudem wird auf Besonderheiten in einzelnen Ländern durch an die Ländernamen angefügte Buchstabenfußnoten (Spalte 2) hingewiesen. Erläuterungen geben Aufschluss über mögliche Einschränkungen, die sich für eine international vergleichende Interpretation der Ergebnisse ergeben (siehe Abschnitt 4.3.4).

In den Spalten 3, 4 und 5 von Tabelle A.4 ist gekennzeichnet, welche Teilnehmer und Benchmark-Teilnehmer von IGLU 2016 auch in 2011, in 2006 beziehungsweise in 2001 beteiligt waren. Die Tabellenspalten 6 bis 9 verweisen auf Besonderheiten bezüglich der Definition und Ausschöpfung der nationalen Zielpopulation, die Tabellenspalten 10 bis 14 auf Besonderheiten bezüglich der *Schul- und Schülerteilnahmequoten*. In Tabellenspalte 15 sind *Besonderheiten bezüglich der Leistungsmessung* dokumentiert.

Das Durchschnittsalter der getesteten Schülerinnen und Schüler ist der sechsten Tabellenspalte zu entnehmen. England, Malta, Neuseeland sowie Trinidad und Tobago wählten als äquivalente nationale Zielpopulation gemäß den Kriterien ‚formale Beschulungszeit' und ‚Durchschnittsalter des Schülerjahrgangs' eine höhere Jahrgangsstufe.

In Spalte 8 wird der *Ausschöpfungsgrad der nationalen Zielpopulation in Prozent bezogen auf die internationale Vorgabe* (100 %) illustriert. Hier zeigt sich für die Teilnehmer Georgien und Kanada, dass die internationale Zielpopulation nicht vollständig durch die nationale Zielpopulation abgedeckt wird. Für Georgien ist der geringere Ausschöpfungsgrad dadurch bedingt, dass hier nur Schülerinnen und Schüler getestet wurden, die in Georgisch und Aserbaidschanisch unterrichtet werden. In Kanada wurden nur Schülerinnen und Schüler aus den Provinzen Alberta, Britisch Kolumbien, Manitoba, Neubraunschweig, Neufundland, Ontario, Québec und Saskatchewan getestet.

In Spalte 9 sind die *Ausschlüsse von der Zielpopulation* als Gesamtausschlussquote in Prozent aufgeführt. Hier zeigt sich, dass die internationale Vorgabe von maximal 5 Prozent von 13 Teilnehmern und drei Benchmark-Teilnehmern sowie von Dänemark (3. Jahrgangsstufe, PIRLS *Literacy*) überschritten wird. Deutschland gehört zu den Teilnehmern, deren Gesamtausschlussquote auf der Ebene der nationalen Zielpopulation unter 5 Prozent liegt (für Details zu Stichprobe und Beteiligungsquoten in Deutschland siehe Abschnitt 7).

Tabellenspalten 10 bis 14 sind die *Schulteilnahmequoten* (ohne und mit Ersatzschulen), die *Schülerteilnahmequoten* sowie die *Gesamtteilnahmequoten* (ohne und mit Ersatzschulen) zu entnehmen. Während die meisten Teilnehmer und Benchmark-Teilnehmer die internationalen Vorgaben erfüllen, zeigt sich für Belgien (Flämische Gemeinschaft), Hongkong, Kanada, die Niederlande, Nordirland, die USA und Südafrika (5. Jahrgangsstufe), dass sie eine *Schulteilnahmequote* von 85 Prozent nur unter Hinzunahme von Ersatzschulen erreichen. Québec gelingt dies auch unter Berücksichtigung der Ersatzschulen nicht. Eine *Schülerteilnahmequote* von 85 Prozent erreichen alle an IGLU 2016 teilnehmenden Bildungssysteme. Eine *kombinierte Schüler- und Schulgesamtteilnahmequote* von 75 Prozent erreichen Hongkong, die Niederlande und die USA hingegen nur unter Hinzunahme von Ersatzschulen. Québec liegt auch mit Ersatzschulen noch unterhalb dieser Vorgabe.

Tabellenspalte 15 dokumentiert den Anteil der Schülerinnen und Schüler ohne skalierbare Leistungswerte im Lesetest bei IGLU 2016. Da diese Information erst seit der Erhebung im Jahr 2011 für IGLU berichtet wird, finden sich entsprechende Angaben nicht in den Tabellen A.1 und A.2. Wie Tabelle A.4 zu entnehmen ist, gibt es bei IGLU 2016 keinen Teilnehmer oder Benchmark-Teilnehmer, der die internationale Vorgabe von maximal 15 Prozent an Schülerinnen und Schülern ohne skalierbare Leistungswerte überschreitet.

4.3.3 Bedeutsamkeit der Ausschlussquote für den internationalen Vergleich

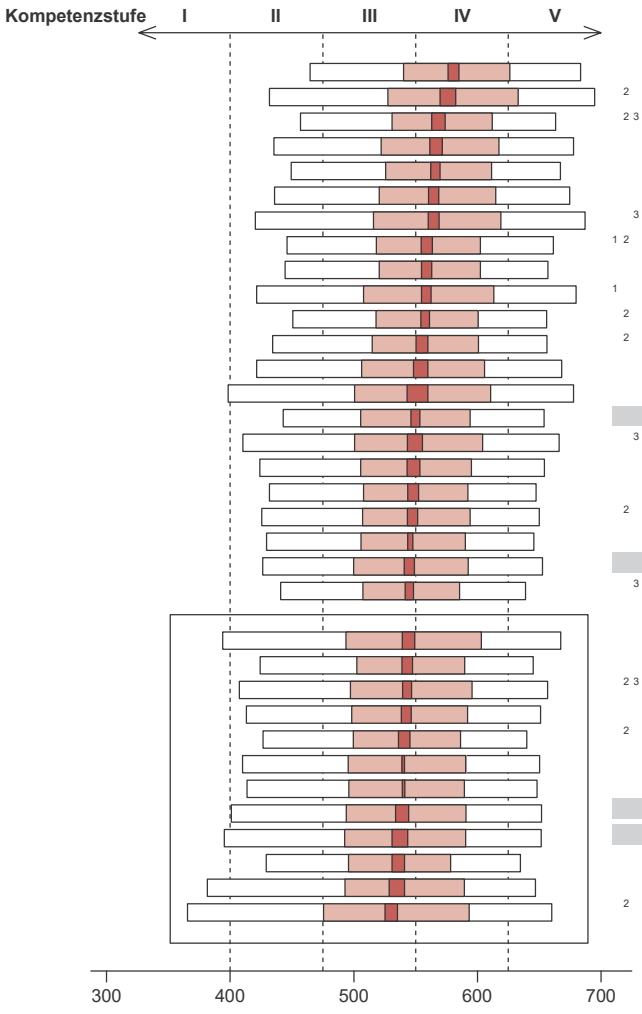
Die *Ausschlüsse von der Zielpopulation* liegen für Deutschland mit einer Gesamtausschlussquote von 4.2 Prozent im Toleranzbereich, das heißt unterhalb von 5 Prozent (siehe Abschnitt 4.3.2). Bei manchen Teilnehmern ist sie deutlich geringer (z. B. in Taiwan mit 0.9 % oder in Finnland mit 2.4 %), andere Teilnehmer wie Singapur (11.1 %), Hongkong (10.1 %), Dänemark (9.8 %), Lettland (7.9 %) oder Österreich (5.6 %) erreichen Ausschlussquoten, die deutlich über der internationalen Vorgabe liegen.

In Deutschland ist die Gesamtausschlussquote im Vergleich zu IGLU 2011 (1.9 %) oder auch IGLU 2006 (0.7 %) deutlich gestiegen. Von den 4.2 Prozent sind 1.4 Prozentpunkte der Ausschlüsse auf Schulebene zu begründen, 2.8 Prozentpunkte auf Schülerebene (2 % Schülerinnen und Schüler, die weniger als ein Jahr in der Testsprache unterrichtet wurden und deren Muttersprache nicht die Testsprache ist, und 0.8 Prozent Kinder, die aus Sicht der Lehrkräfte körperlich, emotional oder aus geistigen Gründen nicht in der Lage waren, den Test selbstständig zu bearbeiten). Die erhöhte Gesamtausschlussquote in 2016 ist damit mit dem erhöhten Anteil von Ausschlussgründen auf Schülerebene zu begründen und hängt mit den, im Zuge von Inklusion und der seit 2015 zu beobachtenden, ansteigenden Zuwanderung nach Deutschland, gestellten Herausforderungen an das Bildungswesen zusammen.

Unterstellte man, dass Ausschlüsse von der Zielpopulation in erster Linie Schülerinnen und Schüler mit eher geringen Kompetenzen betreffen, so wäre zu fragen, welche Mittelwerte die Viertklässlerinnen und Viertklässler in Deutschland bei vergleichbaren Ausschlussquoten erzielt hätten. Die Abbildung 2.4 zeigt Antworten auf diese hypothetische Frage. Dargestellt sind verschiedene Verortungen Deutschlands auf der Gesamtskala Lesen (siehe Kapitel 3, Abbildung 3.6 in diesem Band), die sich ergäben, wenn für Deutschland Ausschlussraten zugrunde lägen, die vergleichbar sind mit denen ausgewählter Teilnehmer, die die Vorgabe von 5 Prozent überschreiten. Als Referenzpunkte sind in der Abbildung die reellen Kompetenzmittelwerte für Deutschland und für diejenigen Teilnehmerstaaten angegeben, die auf der Gesamtskala Lesen signifikant besser als oder nicht signifikant verschieden von Deutschland abschneiden. Der rechten Seite der Abbildung ist zu entnehmen, ob die Leistungsmittelwerte der dargestellten Teilnehmer unter Berücksichtigung der Ausschlussquoten signifikant besser oder schlechter abschneiden als die Leistungsmittelwerte für Deutschland. Symbolisiert sind signifikant höhere Werte durch ein Plus (+), signifikant niedrigere Werte durch ein Minus (-).

Abbildung 2.4 veranschaulicht, dass eine Erhöhung der Ausschlussquoten für Deutschland im Rahmen der internationalen Vorgaben (max. 5 %) keinen bedeutsamen Einfluss auf das Abschneiden der Viertklässlerinnen und Viertklässler in Deutschland im internationalen Vergleich gehabt hätte. Es ist im Umkehrschluss auch zu vermuten, dass die Leistungsmittelwerte der Teilnehmer Singapur, Hongkong, Dänemark oder Lettland bei geringeren Ausschlussquoten unterhalb der für diese Teilnehmer gefundenen Leistungsmittelwerte lägen. Ausschlussquoten in Höhe von 7.9 Prozent (wie in Lettland) oder 11.1 Prozent (wie in Singapur) würden zwar den Leistungsmittelwert der Schülerinnen und Schüler in Deutschland und damit auch die Platzierung Deutschlands im Ländervergleich signifikant positiv beeinflussen. Insgesamt aber bleibt festzuhalten, dass die Ausschlussquoten anderer Teilnehmer nur begrenzt als Begründung für das Abschneiden Deutschlands im internationalen Vergleich herangezogen

Abbildung 2.4: Veränderung des Mittelwerts von Viertklässlerinnen und Viertklässlern in Deutschland beim Vergleich der Leistung auf der Gesamtskala Lesen in Abhängigkeit von der Ausschlussquote



	Teilnehmer	Ausschluss-quote in %	M	(SE)	Signifikanz zu Deutschland bei einer Ausschlussquote in Prozent von 4.2	5.0	7.9	11.1
	Russische Föderation	3.8	581	(2.2)	+	+	+	+
2	*Singapur*	11.1	576	(3.2)	+	+	+	+
2 3	*Hongkong*	10.1	569	(2.7)	+	+	+	+
	Irland	3.1	567	(2.5)	+	+	+	+
	Finnland	2.4	566	(1.8)	+	+	+	+
	Polen	3.9	565	(2.1)	+	+	+	+
3	Nordirland	3.0	565	(2.2)	+	+	+	+
1 2	*Norwegen (5. Jgst.)*	5.3	559	(2.3)	+	+	+	+
	Taiwan	0.9	559	(2.0)	+	+	+	+
1	England	3.7	559	(1.9)	+	+	+	+
2	*Lettland*	7.9	558	(1.7)	+	+	+	+
2	*Schweden*	5.2	555	(2.4)	+	+	+	
	Ungarn	4.5	554	(2.9)	+	+	+	
	Bulgarien	4.3	552	(4.2)	+	+		
	Deutschland[A]	11.1	550	(1.8)	+	+		
3	USA	4.8	549	(3.1)	+	+		
	Litauen	4.2	548	(2.6)	+	+		
	Italien	4.9	548	(2.2)	+	+		
2	*Dänemark*	9.8	547	(2.1)	+	+		
	Macau	3.6	546	(1.0)	+	+		-
	Deutschland[A]	7.9	545	(2.1)	+			
3	Niederlande	3.1	545	(1.7)	+			-
	Australien	4.8	544	(2.5)				
	Tschechien	3.4	543	(2.1)				-
2 3	*Kanada*	7.5	543	(1.8)				-
	Slowenien	2.4	542	(2.0)				-
2	*Österreich*	5.6	541	(2.4)				-
	VG OECD	–	**541**	**(0.5)**			**-**	**-**
	VG EU	–	**540**	**(0.5)**			**-**	**-**
	Deutschland[A]	5.0	539	(2.7)				-
	Deutschland	4.2	537	(3.2)			-	-
	Kasachstan	4.9	536	(2.5)			-	-
	Slowakei	4.8	535	(3.1)			-	-
2	*Israel*	24.9	530	(2.5)		-	-	-

☐ Nicht statistisch signifikant vom deutschen Mittelwert abweichende Staaten ($p > .05$).

Kursiv gesetzt sind die Teilnehmer, für die von einer eingeschränkten Vergleichbarkeit der Ergebnisse ausgegangen werden muss.

\+ = Statistisch signifikant positive Abweichungen ($p < .05$) zu dem für Deutschland ermittelten Leistungsmittelwert unter Berücksichtigung der in der Spaltenüberschrift angegebenen Ausschlussquote.

\- = Statistisch signifikant negative Abweichungen ($p < .05$) zu dem für Deutschland ermittelten Leistungsmittelwert unter Berücksichtigung der in der Spaltenüberschrift angegebenen Ausschlussquote.

1 = Die nationale Zielpopulation entspricht nicht oder nicht ausschließlich der vierten Jahrgangsstufe.

2 = Der Ausschöpfungsgrad und/oder die Ausschlüsse von der nationalen Zielpopulation erfüllen nicht die internationalen Vorgaben.

3 = Die Teilnahmequoten auf Schul- und/oder Schülerebene erreichen nicht die internationalen Vorgaben.

A = Berechnung für Deutschland nach einem höheren Ausschluss von Schülerinnen und Schülern mit geringen Kompetenzwerten.

werden können. Denn sogar unter Ausschluss von 11.1 Prozent der leistungsschwächsten Viertklässlerinnen und Viertklässler aus der deutschen Stichprobe würden etwa Schülerinnen und Schüler in der Russischen Föderation, Singapur, Hongkong, Irland, Finnland oder auch in Polen und Nordirland weiterhin signifikant bessere Leistungsergebnisse im Lesen erzielen als Viertklässlerinnen und Viertklässler in Deutschland.

4.3.4 IGLU 2016 Fußnotensystem zur Klassifikation von Teilnahmebedingungen

Die Besonderheiten der Teilnahmebedingungen sind in den Tabellen A.1, A.2, A.3 und A.4 dokumentiert (siehe Anhang A in diesem Band). Im vorliegenden Berichtsband wurde in Anlehnung an die internationale Berichterstattung für die Darstellung der Ergebnisse ein differenziertes Fußnotensystem entwickelt, das auf die vier Erhebungszyklen von IGLU angewendet werden kann. Die Fußnoten benennen Besonderheiten der Stichproben der einzelnen Teilnehmer. Mit dieser Übersicht ist für eine international-vergleichende Interpretation von Ergebnissen eine Grundlage gegeben, Einschränkungen der Vergleichbarkeit systematisch zu betrachten, die aus Besonderheiten der nationalen Stichproben resultieren. Die nachfolgend aufgeführten Fußnoten werden analog zur Berichtslegung zu IGLU 2011 (Tarelli, Wendt, Bos & Zylowski, 2012) zur Kennzeichnung von Besonderheiten nationaler Stichproben verwendet:

1	=	Die nationale Zielpopulation entspricht nicht oder nicht ausschließlich der vierten Jahrgangsstufe.
2	=	Der Ausschöpfungsgrad und/oder die Ausschlüsse von der nationalen Zielpopulation erfüllen nicht die internationalen Vorgaben.
3	=	Die Teilnahmequoten auf Schul- und/oder Schülerebene erreichen nicht die internationalen Vorgaben.

Die Fußnoten 4, 5 und 6, die im Rahmen der Berichterstattung zu IGLU 2011 (vgl. ebd.), TIMSS 2011 (Wendt, Tarelli, Bos, Frey & Vennemann, 2012) beziehungsweise TIMSS 2015 (Wendt, Bos, Kasper, Walzebug, Goy & Jusufi, 2016) definiert wurden, kommen in der Berichtslegung zu IGLU 2016 nicht vor. Um eine Konsistenz der Fußnotennummerierung in der IGLU- und TIMSS-Berichterstattung zu gewährleisten, werden diese Nummern bei IGLU 2016 nicht anderweitig vergeben.

Für IGLU 2016 wurden im Vergleich zu IGLU 2011 und analog zur Berichterstattung bei TIMSS 2015 zwei weitere Fußnoten (7 und 8) definiert, die zum einen die Teilnahme an PIRLS *Literacy* (Fußnote 7), zum anderen Besonderheiten im Hinblick auf Trendvergleiche mit den Zyklen IGLU 2001, IGLU 2006 und IGLU 2011 (Fußnote 8) betreffen (siehe Anhang A.4 in diesem Band).

Kursivschreibung der Ländernamen und Kennzahlen

Ebenfalls analog zur Berichtslegung von IGLU 2011 (Tarelli et al., 2012) wird neben Fußnoten durch Kursivschreibung der Staatennamen auf eine eingeschränkte Vergleichbarkeit hingewiesen. Bei IGLU 2016 erfolgte dies, wenn für Teilnehmer mindestens eins der beiden Kriterien zutrifft:

- die Gesamtausschlussquote liegt über der internationalen Vorgabe von 5 Prozent (siehe Tabelle A.4 in diesem Band; Spalte 9; *Ausschlüsse von der nationalen Zielpopulation (Gesamtquote) in %*),
- die Schüler- und Schulgesamtteilnahmequote (mit Ersatzschulen) liegt unter 75 Prozent (siehe Tabelle A.4 in diesem Band; Spalte 14; *Gesamtteilnahmequote in %*),

Kursive Formatierungen der Namen der Teilnehmer und Benchmark-Teilnehmer in den Abbildungen und Tabellen in diesem Band werden durch folgende Fußnote erläutert: „Kursiv gesetzt sind die Teilnehmer, für die von einer eingeschränkten Vergleichbarkeit der Ergebnisse ausgegangen werden muss“. In Trenddarstellungen weisen entsprechende Kursivsetzungen darauf hin, dass in den betreffenden Bildungssystemen Besonderheiten in den Erhebungsbedingungen vorliegen und Interpretationen von Ergebnissen im Vergleich der Studienzyklen nur unter deren Berücksichtigung erfolgen sollten (siehe hierzu auch die Erläuterungen zu Fußnote 8 sowie die Tabellen A.1 bis A.4 im Anhang A in diesem Band).

5 Entwicklung und Charakteristika der Instrumente

In IGLU werden zur Erfassung der Lesekompetenzen Leistungstests eingesetzt. Die für das schulische Lernen relevanten Hintergrundinformationen werden mit Hilfe von Fragebögen auf System-, Schul-, Lehrer-, Schüler- und Elternebene umfassend erhoben.

5.1 Leistungstests

Die Leistungstests, die in IGLU eingesetzt werden, dienen der Ermittlung von Leistungsständen der Schülerinnen und Schüler im Bereich des Leseverstehens. Die so ermittelten Schülerleistungen erlauben, Aussagen im internationalen Vergleich über Leistungsunterschiede zwischen gesamten Schülerpopulationen und zwischen spezifischen Schülergruppen für die übergreifenden Kompetenzskalen und die Subdomänen zu treffen (siehe Abschnitt 4.1). Die Leistungen sollen für die übergreifende Gesamtskala zur Messung der Lesekompetenz und für die dem Rahmenkonzept entsprechenden Subskalen auswertbar sein (siehe Kapitel 3 in diesem Band).

Die in IGLU eingesetzten Leistungstests bestehen aus verschiedenen Erzähl- und Sachtexten und den dazu entwickelten Verständnisaufgaben (Martin et al., 2015; Mullis et al., 2015). In der Regel werden ungefähr 60 Prozent der Aufgabenblöcke zum Zwecke der Verlinkung mit den zurückliegenden Erhebungszyklen aus den vorangegangenen Erhebungszyklen übernommen. Dabei stammen etwa 75 Prozent aus den letzten und knapp 25 Prozent aus dem jeweils vorletzten Zyklus. Etwa 40 Prozent der Blöcke werden für einen Erhebungszyklus neu entwickelt. Die Auswahl der Texte und die Entwicklung der Testaufgaben erfolgen in einem kooperativen Prozess, an dem Vertreterinnen und Vertreter der Teilnehmerstaaten, die internationale Studienleitung und eine Gruppe aus ausgewählten internationalen Expertinnen und Experten (der sogenannten *Reading Development Group*) beteiligt sind. Aufgaben können aus jedem Teilnehmerstaat vorgeschlagen und eingereicht werden. Nach der Einreichung

erfolgt eine umfassende Begutachtung. Die *Reading Development Group* wählt aus den Vorschlägen jenes Textmaterial aus, das authentischen Leseerfahrungen von Viertklässlerinnen und Viertklässlern in unterschiedlichen Kulturen entspricht und den Anforderung der Messung von Lesekompetenzen im Hinblick auf den Umfang und den Inhalt gerecht wird. Geprüft wird das Material unter anderem hinsichtlich der Angemessenheit der Themen, der Informationsdichte und des Niveaus darin auftretender linguistischer Charakteristika. Besonderes Augenmerk wird zudem auf Aspekte geschlechtsspezifischer, kultureller und religiöser Fairness der Texte gelegt. Texte, die den Prozess der Begutachtung erfolgreich durchlaufen haben, werden der Gruppe der *National Research Coordinators* vorgestellt. Diese Gruppe prüft erneut die Inhalte für die teilnehmenden Staaten und Regionen. Nach Bestimmung der Texte werden – ebenfalls in Kooperation mit den Teilnehmerstaaten – Testaufgaben zu den Texten konzipiert. Verbindlich definierte Kriterien, die in einem Leitfaden formuliert sind, gewährleisten, dass die Aufgaben die in der Rahmenkonzeption festgelegten Inhalte und Anforderungen angemessen repräsentieren (Martin et al., 2017; Mullis & Prendergast, 2017).

Diesem Prozess folgen im Rahmen des Feldtests ein Jahr vor der Haupterhebung die Erprobung der Textauswahl und die Überprüfung der empirischen Qualität der Aufgaben. Von insgesamt 12 im Feldtest neu eingesetzten Texten wurden sechs Texte in der Hauptuntersuchung eingesetzt (Mullis & Prendergast, 2017). Ausgewählt wurden diejenigen Texte, deren Testaufgaben für die Schülerinnen und Schüler aller Teilnehmerstaaten vergleichbare Messeigenschaften aufweisen konnten.

5.1.1 Charakteristika der Testaufgaben

In IGLU 2016 wurden 12 Lesetexte und entsprechende Aufgaben zu den Texten eingesetzt, davon sechs als informierende und sechs als erzählende Lesetexte. Vier von diesen 12 Texten wurden neu ergänzt, die anderen acht stammen aus den Erhebungszyklen 2001, 2006 und 2011 (Mullis & Prendergast, 2017). Zwei der vier neu in IGLU 2016 eingesetzten Texte wurden aus der ergänzenden Studienkomponente PIRLS *Literacy* (siehe Abschnitt 4.2.1 in diesem Kapitel) in den Aufgabenpool übernommen, um eine Vergleichbarkeit der Testleistungen in IGLU und PIRLS *Literacy* herzustellen (ebd.). Dabei handelt es sich um einen Sach- und einen Erzähltext in leichterer Sprache (siehe auch Kapitel 3, Tabelle 3.1 in diesem Band). Durch den Einsatz von in den Erhebungszyklen 2001, 2006 oder 2011 verwendeter und neuer Texte findet eine Verknüpfung der vergangenen Erhebungszyklen mit dem aktuellen statt, so dass eine Einschätzung der Veränderung in den Leistungsergebnissen im Trend ermöglicht ist.

Alle Lesetexte sind kindgerecht aufbereitet, ansprechend illustriert und sprachlich klar formuliert (siehe die beiden Beispieltexte in Kapitel 3, Abschnitte 4.2.1 und 4.2.2 in diesem Band). An jeden Text schließen sich 12 bis 17 Aufgaben an, die in einem geschlossenen (*Multiple Choice*) oder offenen Antwortformat erstellt sind und auf unterschiedlichen Schwierigkeitsniveaus verschiedene Aspekte des Leseverständnisses erfragen. Bei geschlossenen Formaten wählen Schülerinnen und Schüler aus vier vorgegebenen Antworten die richtige Lösung aus. Bei offenen Formaten ist von den Schülerinnen und Schülern verlangt, ihre Antworten in einem Textfeld frei zu formulieren. Die Auswertung der Aufgaben erfolgt nach international einheitlichen Auswertungskriterien.

5.1.2 Rotation der Testaufgaben

Jede Schülerin beziehungsweise jeder Schüler bearbeitet während der Testsitzung ein Testheft. Darin enthalten sind zwei Lesetexte mit dazugehörigen Aufgaben.

Wie in den meisten internationalen Leistungsstudien, liegt auch der Testkonzeption in IGLU ein multiples Matrixdesign zugrunde, das eine Rotation der 12 Texte (6 Erzähl- und 6 Sachtexte) über 16 unterschiedliche Testhefte vorsieht (siehe Tabelle 2.3). Das Grundprinzip dieses Designs sieht vor, dass nicht allen Schülerinnen und Schülern sämtliche Aufgaben vorgelegt werden, sondern jede Schülerin und jeder Schüler lediglich einen Teil der Aufgaben bearbeitet. Eine Schülerin beziehungsweise ein Schüler bearbeitet jeweils zwei ausgewählte Lesetexte (einen Erzähl- und einen Sachtext), die in einem Testheft arrangiert sind.

Die in IGLU gezogene Stichprobe ist ausreichend groß, so dass mit Hilfe statistischer Methoden Leistungswerte dank des multiplen Matrixdesigns für die Population präzise geschätzt werden können. Dieses Testdesign bietet den Vorteil, auch die Leistung von Schülerinnen oder Schülern vergleichen zu können, die unterschiedliche Testteile bearbeitet haben.

Tabelle 2.3: Testheftdesign in IGLU 2016

	Testheft															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Geschichten-heft
Teil 1	Inf 1	Lit 2	Inf 2	Lit 2	Inf 2	Lit 4	Inf 4	Lit 5	Inf 5	Lit 5	Lit 2	Inf 2	Inf 4	Lit 5	Inf 5	Lit 6
Teil 2	Lit 1	Inf 1	Lit 2	Inf 3	Lit 3	Inf 3	Lit 4	Inf 4	Lit 1	Inf 5	Inf 3	Lit 4	Lit 1	Inf 1	Lit 3	Inf 6

☐ Lesetexte aus PIRLS *Literacy*.

IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

5.1.3 Freigegebene Aufgaben

Informationen zu der Art und den Inhalten der in IGLU eingesetzten Lesetexte und Aufgaben werden nach der Berichtslegung der Öffentlichkeit zugänglich gemacht. Die internationale Studienleitung wird im Frühjahr 2018 Lesetexte und Aufgabenblöcke der IGLU 2016-Testhefte veröffentlichen. Die in diesem Bericht verwendeten Aufgabenbeispiele entstammen dieser Auswahl (siehe Kapitel 3 in diesem Band). In den letzten Jahren hat die IEA ihr Verfahren der Nutzung von Testaufgaben verändert. Eine Nutzung von Aufgaben muss bei der IEA beantragt werden.[1]

1 Weiterführende Informationen sind unter folgendem Link verfügbar: http://www.iea.nl/permissions.html

5.1.4 Ergänzende Aufgaben zum Leseverständnis auf Wort- und Satzebene

In Deutschland wurden im Rahmen von IGLU 2016 ergänzend zu den international vorgegebenen IGLU-Tests zwei Testkomponenten des *Leseverständnistests für Erst- bis Siebtklässler, Version II* (ELFE II) eingesetzt (Lenhard, Lenhard & Schneider, 2017). Der ELFE II dient der Erfassung der Leseflüssigkeit, der Lesegenauigkeit und des Leseverständnisses auf Wort-, Satz- und Textebene und differenziert nach dem Kompetenzniveau von Schülerinnen und Schülern vom Ende der ersten bis zum Beginn der siebten Klassenstufe. Es handelt sich um einen standardisierten Test, der sich bewährt hat, um zum Beispiel Defizite im Schriftspracherwerb frühzeitig zu erkennen und gegebenenfalls unterstützende Maßnahmen einleiten zu können.

In IGLU 2016 wurden die ELFE II-Untertests zum Wort- und Satzverständnis eingesetzt, nicht aber zum Textverständnis. Beide Tests wurden im Anschluss an den IGLU-Test als *Papier-und-Bleistift*-Tests von den Schülerinnen und Schülern bearbeitet (siehe Abschnitt 6.1). Bei dem Wortverständnistest (75 Items) geht es darum, passend zu einem dargestellten Bild eines von vier möglichen Wörtern zu erkennen und anzukreuzen. Ist beispielsweise eine Ente abgebildet, so müssen Schülerinnen und Schüler aus den vier Antwortalternativen „Ente“ – „Stein“ – „Blume“ – „Fisch“ das zum Bild passende Wort erkennen und ankreuzen. Beim Satzverständnistest (36 Items) hingegen ist aus fünf Antwortalternativen dasjenige Wort zu unterstreichen, das den vorgegebenen Satz sinnvoll ergänzt. Ein Beispiel hierzu wäre der Satz „Mit einem Füller – Bein – Kuchen – Kopf – Hals kann man schreiben.“

Der Test ist sowohl in der Einzeldiagnostik als auch in der Testung großer Gruppen und Stichproben verwendbar. Die Durchführung erfolgte nach Manual standardisiert. Für den Wortverständnistest und für den Satzverständnistest hatten die Schülerinnen und Schüler jeweils 3 Minuten Zeit. Für jede richtig gelöste Aufgabe wurde ein Punkt gegeben. Nicht erreichte oder ausgelassene Aufgaben sind als falsch gewertet worden.

Die *odds-even Split-Half*-Reliabilität erreicht mit .99 für den Wortverständnistest und .98 für den Satzverständnistest ähnliche hohe Werte wie sie in der Normstichprobe festgestellt worden sind (.98 für den Wortverständnistest und .94 für den Satzverständnistest). Da für den ELFE II schulmonatsgenaue Normwerttabellen vorliegen, wurde tagesgenau geprüft, in welchem Schulmonat der vierten Klassenstufe sich die Schülerinnen und Schüler befunden haben. 61.8 Prozent der Schülerinnen und Schüler befanden ich im Bereich von sieben bis acht Schulmonaten und 38.2 Prozent im Bereich von neun bis zehn Schulmonaten. Beide Normwerttabellen wurden angewendet und verglichen. Dabei wurden keine statistisch bedeutsamen Differenzen in logistischen Regressionen und Varianzvergleichen festgestellt, die auf eine Notwendigkeit der Anwendung unterschiedlicher Normwerttabellen hindeuten (Wortverständnistest: *beta stand* = -.48; *t* = -0.48; Satzverständnistest: *beta stand* = -.88; *t* = -0.90). Da eine größere Zahl von Schülerinnen und Schülern sieben bis acht Schulmonate in der vierten Klassenstufe aufwiesen, wurden in der Folge die entsprechenden Normwerttabellen für beide Testteile verwendet.

Abschließend wurden die Ergebnisse der Testteile kombiniert, um diagnostische Schwellenwerte abzuleiten. Dafür wurden, wenn gültige Werte für beide Testteile vorlagen, die normierten Werte für das Wortverständnis und das Satz-

verständnis summiert und anhand einer Normtabelle für die ELFE-II-Kurzform erneut in normierte Werte umgewandelt.

Die beiden Untertests aus ELFE II wurden in IGLU 2016 mit dem Ziel eingesetzt, insbesondere die Leistungen der leseschwachen Schülerinnen und Schüler (Kompetenzstufen I und II) differenzierter erfassen zu können als in den Erhebungszyklen zuvor, da über die Leistungen dieser Kinder bislang kaum Aussagen getroffen werden konnten. Mit dem Lesen auf Wortebene sind hierarchieniedrige Prozesse des Lesens angesprochen, die eine wichtige Grundlage für das Beherrschen hierarchiehöherer Prozesse (= Lesen auf Satz- und Textebene) darstellen (z. B. Christmann & Groeben, 2001). Die Befunde der in IGLU 2016 eingesetzten Untertests aus ELFE II liefern damit differenzierte Informationen zu dem Leseverstehen von Schülerinnen und Schülern auf der Schwelle von hierarchieniedrigen zu hierarchiehöheren Leseprozessen. Befunde hierzu werden in den Kapiteln 3 und 10 im vorliegenden Band vorgestellt.

5.2 Kontextfragebögen

Mit Hilfe von schriftlichen Befragungen der an IGLU teilnehmenden Schülerinnen und Schüler und der an ihrem Lernprozess beteiligten Akteurinnen und Akteure können Hintergrundinformationen erhoben werden, die mit schulischem Lernen – insbesondere in Bezug auf den Erwerb von Lesekompetenzen bezogen – im Zusammenhang stehen können (Hooper & Fishbein, 2017). Wie schon für die Leistungstests beschrieben (siehe Abschnitt 5.1), erfolgt auch die Entwicklung der Fragebogenitems in einem kooperativen Prozess der IGLU-Teilnehmerstaaten. Vorschläge für Fragen zur Erfassung von Hintergrundmerkmalen können aus allen Teilnehmerstaaten eingereicht werden. Nach der Einreichung erfolgt eine Begutachtung durch Expertinnen und Experten. Zu jedem Studienzyklus werden neue Fragebogenitems entwickelt und eingereicht. Dies geschieht etwa dann, wenn mit aus den Fragebögen gewonnenen Informationen Unterschiede in schulischen Leistungen von Schülerinnen und Schülern verschiedener Teilnehmerstaaten und Regionen umfassender als zuvor erfasst und damit interpretiert werden können. Maßgeblich ist die inhaltliche Nähe zu und der Erkenntnisgewinn mit den durch die Rahmenkonzeption fokussierten Inhalten und Anforderungen (siehe Abschnitt 4.1). Sichergestellt wird auch, dass die ausgewählten Fragen die Testziele der IGLU-Rahmenkonzeption repräsentieren und sich in allen Teilnehmerstaaten einsetzen lassen. Die Überprüfung der Qualität erfolgt im Feldtest. Die von der internationalen Studienleitung bestimmten Inhalte der Kontextfragebögen sind für alle Teilnehmerstaaten verbindlich (Mullis & Martin, 2015). Da sich diese Inhalte nicht zwingend mit Fragen decken, die auch Wissenschaft und Politik in einzelnen Staaten vorrangig beschäftigen, besteht die Möglichkeit der nationalen Erweiterung von Fragebögen. So können Fragebogenitems national ergänzend zu den international verbindlich vorgegebenen Fragen in einem Staat eingesetzt werden. In Deutschland wurden der Schüler-, Eltern-, Lehrkräfte- und der Schulfragebogen um nationale Ergänzungen erweitert.

5.2.1 Datenschutzrechtliche Begutachtung der Kontextfragebögen

Die in IGLU 2016 implementierten Verfahren der Datenerhebung und die Instrumente wurden durch die Kultusministerien der Länder datenschutzrechtlich geprüft und anschließend – teilweise mit kleineren länderspezifischen Anpassungen – für die Erhebung zugelassen.

5.2.2 Inhalte der Kontextfragebögen

Die in IGLU 2016 eingesetzten Fragebögen thematisieren unter anderem die nachfolgend aufgeführten Aspekte (Hußmann, Wendt, Bos & Rieser, 2018).

Der *Schülerfragebogen* umfasst Fragen zum soziodemografischen und soziokulturellen Hintergrund des Kindes (z. B. Alter, Geschlecht, Sprachgebrauch zu Hause, Muttersprache) und Fragen zum Leseverhalten, zur Lesemotivation und zum Leseselbstkonzept. Erfragt werden die Teilnahme an außerunterrichtlichen Angeboten (Ganztags- und Förderangeboten) und Freizeitaktivitäten.

Informationen, die aus dem *Elternfragebogen* gewonnen werden, ergänzen die Angaben der Schülerinnen und Schüler zur familiären Lernumwelt. Darin erfragte Themen sind unter anderem die einer Familie zur Verfügung stehenden Ressourcen, soziodemographische Daten der Eltern (z. B. höchster Bildungsabschluss, Fragen zur Berufstätigkeit, Zusammensetzung der Familie), der Bildungshintergrund der Eltern, Bildungsaspirationen der Eltern, ebenso wie frühe Lernerfahrungen des Kindes sowie lernunterstützende Aktivitäten und Fragen zum kognitiven Anregungsgehalt der familiären Lernumwelt.

Der *Fragebogen für die Deutschlehrkräfte* dient der systematischen Erfassung des Unterrichts und spezifischer Klassenmerkmale. Neben Personenmerkmalen (Alter, Geschlecht, Lehrerfahrung) sind Fragen zur Ausbildung und Qualifizierung, zu Merkmalen und zur Ausstattung der Klasse, zu Einstellungen unterrichtsrelevanter Aspekte, Unterrichtspraxis und Lehrmethoden, lernrelevanten Voraussetzungen der Schülerinnen und Schüler sowie persönlichen Einstellungen zum Lesen zu beantworten. Zusätzliche Informationen wie Schullaufbahnempfehlungen, Förderbedarfe oder Schulnoten werden mit einer *Schülerteilnahmeliste* erhoben, die ebenfalls durch eine Lehrkraft ausgefüllt wird.

Mit dem von der Schulleitung auszufüllenden *Schulfragebogen* werden Informationen zu Schulcharakteristika (Schulgröße, Lehrkräfte, Unterrichtszeiten, soziale und geographische Lage), zur Schulorganisation, zum Kontext und zur Ausstattung einer Schule, pädagogischen Zielsetzungen, zu Curricula, Ressourcen in Form von technischer und materieller Ausstattung, zum sozialen Klima, zu Kooperationsstrukturen, Einbindung der Eltern, außercurricularen Aktivitäten, zur Rolle der Schulleitung, zum Umgang mit Heterogenität, zur Ganztagsbetreuung sowie zu Angeboten zum Lesen gewonnen.

Ergänzend zu diesen Fragebögen enthält der *Curriculumfragebogen* Fragen dazu, inwieweit die Curricula der teilnehmenden Länder und Regionen durch die Inhalte der IGLU-Tests abgebildet werden. Ausgefüllt wird der Fragebogen, wie in Abschnitt 4.1.2 erläutert, von Fachexpertinnen und Fachexperten der Staaten und Regionen. Fragen darin umfassen Rahmenbedingungen der Umsetzung des Curriculums, zum Beispiel wer über das jeweilige Curriculum entscheidet, ob und in welcher Form die Implementation des Curriculums evaluiert wird oder ob und, wenn ja, inwiefern Lehrkräfte bei der Umsetzung des Curriculums

unterstützt werden. Die Ergebnisse dieser Befragung gehen in die *Enzyklopädie* (Mullis et al., 2017; Wendt, Walzebug, Bos, Smith & Bremerich-Vos, 2017) ein.

5.2.3 Übersetzung und Gestaltung der Testinstrumente

Die internationale Studienleitung stellt zur Durchführung der Studie jedem Teilnehmerstaat die Test- und Befragungsinstrumente in englischer Sprache bereit. An Schulen eingesetzt werden die Tests und Fragebögen in der Sprache des Teilnehmerlandes beziehungsweise der Teilnehmerregion. Jedes nationale Forschungsteam trägt Verantwortung dafür, die Testinstrumente und Fragebögen in den für den nationalen Kontext relevanten Unterrichtssprachen zu erstellen (Martin et al., 2015; Mullis & Prendergast, 2017). Um ein hohes Maß an Qualität und die Gestaltung nach einheitlichen Standards zu gewährleisten, ist international präzise festgelegt, nach welchen Richtlinien und Abläufen die Übersetzung und die Gestaltung der Instrumente erfolgen sollen. Bei der Übersetzung der Testaufgaben oder Fragen in die jeweilige(n) Unterrichtssprache(n) muss die internationale Vergleichbarkeit gewahrt bleiben. Im Speziellen gilt für die Leistungstests, dass sich durch die Übersetzung weder eine Aufgabenstellung noch eine Antwortoption in ihrem Schwierigkeitsgrad von dem international vorgegebenen Original unterscheiden darf. Hingegen ist bei der Übersetzung der international vorgegebenen Fragen in den Kontextfragebögen zu prüfen, ob sie auch im nationalen Kontext einen Sinn ergeben. Gegebenenfalls ist auf Besonderheiten des nationalen Schulsystems Rücksicht zu nehmen, so dass es in Absprache mit der internationalen Studienleitung zu nationalen Anpassungen kommen kann.

Die Einhaltung der Prozeduren sowie die Qualität der Übersetzung werden von der internationalen Studienleitung sorgfältig kontrolliert. Nationale Anpassungen der international vorgegebenen Instrumente sind nur zulässig, sofern sie durch kulturelle Unterschiede begründet werden können (Yu & Ebbs, 2011). Die Qualität der Übersetzung in andere Sprachen und mögliche nationale Adaptionen werden von Übersetzerinnen und Übersetzern der IEA durch Rückübersetzung ins Englische kritisch überprüft. Dabei identifizierte Abweichungen von Inhalt und Layout des englischen Originals werden in Abstimmung mit der nationalen Studienleitung korrigiert und zulässige nationale Abweichungen dokumentiert.

In Deutschland erfolgte die Übersetzung der Testaufgaben und Fragebögen wie auch im vergangenen Studienzyklus zu IGLU 2011 in Kooperation mit dem IGLU-Forschungsteam aus Österreich. Umgesetzt wurde sie von einer professionellen Übersetzerin, die auf eine breite Erfahrung im sensiblen Umgang mit Begriffen und Formulierungen im Rahmen von Schulleistungsstudien zurückgreifen kann und bereits bei IGLU 2011 diese Prozesse begleitet hat. Im Anschluss an die Übersetzung wurden die Testinstrumente und Fragebögen erneut von Expertinnen und Experten kritisch begutachtet und – sofern sinnvoll – mit den Übersetzungen aus den vorangegangenen Studienzyklen abgeglichen, gegebenenfalls hatte dies kleinere sprachliche Anpassungen zur Folge.

Verzichtet wurde, anders als noch in IGLU 2011, auf die Übersetzung des Elternanschreibens in verschiedene Sprachen. Es lag also nur in deutscher Sprache vor. Das Elternanschreiben informiert die Eltern der teilnehmenden Schülerinnen und Schüler über die Studie, die Organisation der Studien, Komponenten der Leistungsmessung, Inhalte der Befragung und datenschutz-

rechtliche Grundlagen. Teil dieses Elternanschreibens ist auch die schriftliche Bestätigung des Einverständnisses zur Teilnahme des Kindes.

6 Erhebung

Die Haupterhebung von IGLU 2016 fand in Deutschland in der Zeit vom 2. Mai bis zum 3. Juni 2016 statt.

6.1 Aufbau der Untersuchung

Der international vorgegebene Testablauf stellt sicher, dass alle teilnehmenden Schülerinnen und Schüler genügend Zeit zur Bearbeitung der Testaufgaben und zur Beantwortung der Fragen des Schülerfragebogens haben. In Deutschland fanden die Testung und die Befragung der Schülerinnen und Schüler an einem Vormittag statt. Die Schülerinnen und Schüler lösten am Testtag zuerst die international vorgegebenen IGLU-Testaufgaben (siehe Abschnitt 5.1). Zwischen dem ersten und dem zweiten Teil des Testhefts fand eine 10-minütige Pause statt. Für die Bearbeitung beider Testteile waren je 40 Minuten vorgesehen. Dies ergab eine Gesamttestzeit von 80 Minuten (siehe Tabelle 2.4).

Ergänzend wurden in Deutschland im Anschluss an den IGLU-Test kleinere Tests in allen IGLU-Klassen durchgeführt. Diese umfassen mit dem Kurztest zu figuralem Denken einen Teilbereich kognitiver Fähigkeiten (KFT, N2, Heller &

Tabelle 2.4: Untersuchungsablauf von IGLU 2016

Beginn der Testsitzung: Verteilung des Materials, Begrüßung, allgemeine Einweisung	ca. 10 min.
Bearbeitung Lesetest Teil I	**40 min.**
Pause	10 min.
Einweisung in den Lesetest Teil II	ca. 5 min.
Bearbeitung Lesetest Teil II	**40 min.**
Pause [Einsammeln der Testhefte und Austeilen der Schüler- und Elternfragebogen]	20 min.
Einweisung in die Bearbeitung des kognitiven Fähigkeitstests (KFT, N2)	ca. 5 min.
Bearbeitung des kognitiven Fähigkeitstests (KFT, N2)	**7 min.**
Einweisung in die Bearbeitung des Wortverständnistests (ELFE II)	ca. 4 min.
Bearbeitung des Wortverständnistests (ELFE II)	**3 min.**
Einweisung in die Bearbeitung des Satzverständnistests (ELFE II)	ca. 2.5 min.
Bearbeitung des Satzverständnistests (ELFE II)	**3 min.**
Einweisung in den Schülerfragebogen	ca. 5 min.
Bearbeitung des Schülerfragebogens	**ca. 45 min.**
Beenden der Testsitzung: Einsammeln der Materialien	ca. 5 min.
Reine Bearbeitungszeit	**ca. 138 min.**
Gesamtzeit	ca. 204.5 min.

IEA: Progress in International Reading Literacy Study (PIRLS)

Perleth, 2000) sowie Leseverstehen auf Wort- und Satzebene (ELFE II, Lenhard et al., 2017), wie in Abschnitt 5.1.4 erläutert. Die Bearbeitungszeit für den KFT-Subtest betrug 7 Minuten, für die ELFE-Untertests je 3 Minuten.

Der Testtag endete für die teilnehmenden Schülerinnen und Schüler mit der Bearbeitung des Schülerfragebogens. Für die Bearbeitung des Schülerfragebogens hatten die Schülerinnen und Schüler 45 Minuten Zeit.

6.2 Durchführung der Erhebung

Die Datenerhebung wurde, wie auch in allen vorangegangenen Erhebungsrunden, in Deutschland durch die *IEA Hamburg* durchgeführt. Die Erhebung folgte hoch standardisiert und kontrolliert den internationalen Vorgaben entsprechend. Die Arbeit der *IEA Hamburg* umfasste den Kontakt mit den Schulen und die Auswahl der Testleiterinnen und Testleiter, die zumeist Lehramts-, Erziehungswissenschafts- oder Psychologiestudierende höherer Semester waren, und über Erfahrungen im Bereich der Testdurchführung verfügten. Sämtliche Testleiterinnen und Testleiter nahmen vor der Erhebungsphase an einer verbindlichen Schulung zur Testdurchführung teil und erhielten ein schriftliches Manual mit genauen Vorgaben zur Durchführung des Testtages. Sie waren dazu angehalten, den international vorgegebenen Testablauf akkurat einzuhalten. Nach standardisierter Vorgabe wurde den Schülerinnen und Schülern am Testtag jeder Testteil ausführlich erklärt und die Durchführung anhand von Beispielen erläutert. Die Kinder hatten jederzeit die Möglichkeit, Verständnisfragen zur Testbearbeitung zu stellen, die nicht auf den Inhalt bezogen waren.

6.3 Qualitätssicherung

Um verlässliche Aussagen aus international-vergleichenden Schulleistungsstudien wie IGLU ableiten zu können, ist es notwendig, dass der Leistungstest in allen teilnehmenden Staaten unter vergleichbaren Voraussetzungen durchgeführt wird. Mit der Teilnahme an IGLU verpflichten sich alle Staaten zur Einhaltung aller Vorgaben und zur Implementation umfassender qualitätssichernder Maßnahmen.

Unabhängige Expertinnen und Experten wurden von der IEA beauftragt, in allen Teilnehmerstaaten stichprobenartig die Einhaltung der internationalen Vorgaben zu überprüfen. In Deutschland führte das sogenannte *International Quality Control Monitoring* (IQCM) Dr. Kristina Frey vom Institut für Erziehungswissenschaft der Universität Münster durch. Sie begutachtete in einer international vorgegebenen Stichprobe von 10 Prozent sämtlicher Testklassen den Ablauf der Testdurchführung. Beobachtungen wurden in von der internationalen Studienleitung vorgegebene Bögen dokumentiert und mündliche Befragungen mit Schulkoordinatorinnen, Schulkoordinatoren und Testleitungen im Anschluss an die besuchten Testungen durchgeführt. Bei der Qualitätskontrolle in Deutschland wurden keine Mängel in der Einhaltung der Erhebungsstandards festgestellt.

Neben dem IQCM wurde ein von der nationalen Studienleitung verantwortetes Qualitätsmonitoring, das sogenannte *National Quality Control Monitoring* (NQCM), an weiteren 20 zufällig ausgewählten Testschulen durch Mitarbeiterinnen des IFS durchgeführt. Es fanden Qualitätsbeobachtungen am Testvormittag statt, die um Interviews mit der Schulkoordination zur Testqualität, Organisation und zum Belastungsempfinden der Schulen ergänzt und anschlie-

ßend umfassend dokumentiert wurden. Weder die Durchführung des NQCM noch die des IQCM konnten Mängel in der Einhaltung der Erhebungsstandards in Deutschland feststellen.

7 Stichprobe und Beteiligungsquoten in Deutschland

Die für Deutschland zufällig gezogene Schulstichprobe umfasst 208 Schulen, aus allen Ländern der Bundesrepublik. Das *Sampling* ist umfassend im technischen Report dokumentiert (Martin et al., 2017). Die Stichprobenziehung erfolgte gemeinsam und analog zur Stichprobenziehung für TIMSS 2015 (vgl. Wendt et al., 2016). Für die Testung wurde an jeder Schule eine vierte Klasse gezogen. An allen 208 getesteten Schulen konnte die Leistungsmessung planmäßig durchgeführt werden.

Von den 4277 Schülerinnen und Schülern der Stichprobe nahmen 3959 Schülerinnen und Schüler tatsächlich am IGLU-Test teil. Die Nicht-Teilnahme von 318 Schülerinnen und Schülern ist auf verschiedene Gründe zurückzuführen (o.g. Ausschlussgründe, Verlassen der Schule vor dem Testtag oder Abwesenheit am Testtag). Deutschland erreicht damit, ähnlich wie in IGLU 2011, eine Gesamtteilnahmequote (für Schüler *und* Schulen) von 93 Prozent exklusive, beziehungsweise 95 Prozent inklusive Ersatzschulen (siehe Tabelle A.4 im Anhang A in diesem Band). Auf erfreulich hohe Akzeptanz stieß IGLU auch bei Lehrkräften und Schulleitungen. Aus jeder IGLU-Klasse liegt mindestens ein Lehrerfragebogen vor, und die Rücklaufquote der IGLU-Schulfragebögen beträgt für Deutschland 93 Prozent. Die Rücklaufquote der Elternfragebögen fiel in IGLU 2016 allerdings mit 72 Prozent erkennbar kleiner aus als etwa in IGLU 2011 (80%).

In Tabelle 2.5 sind die Rücklaufquoten sowie zentrale Hintergrundmerkmale der Schülerinnen und Schüler im Vergleich von 2001, 2006 und 2011 zu 2016 dargestellt. In Bezug auf die Komposition der Schülerschaft zeigt sich, dass sich der Anteil an Mädchen in der IGLU-Stichprobe und das durchschnittliche Alter praktisch nicht geändert hat. Im Vergleich zu den vorangegangenen Erhebungsrunden ist der Anteil von Kindern mit Migrationshintergrund (beide Eltern im Ausland geboren) in IGLU 2016 mit 19 Prozent statistisch signifikant höher, spiegelt aber auch reale Veränderungen in der Komposition der Schülerschaft an Grundschulen in Deutschland wider (Autorengruppe Bildungsberichterstattung, 2016; Statistisches Bundesamt, 2017). Gleiches gilt für den ebenfalls im Vergleich zu 2011 statistisch signifikanten höheren Anteil an Schülerinnen und Schüler mit besonderen Unterstützungsbedarfen, dieser liegt in IGLU 2016 bei rund 7 Prozent.

Tabelle 2.5: Beteiligungszahlen, Rücklaufquoten und zentrale Charakteristika zu IGLU 2001, 2006, 2011 und 2016

	IGLU 2001	IGLU 2006	IGLU 2011	IGLU 2016
Schulen				
Anzahl	211	405	198	208
Beteiligung am Schulfragebogen (%)	95	96	96	93
Klassen				
Anzahl	393	405	198	208
Lehrkräfte				
Anzahl	393	418	222	227
Beteiligung am Lehrerfragebogen (%)	92	96	96	93
Schülerinnen und Schüler				
Anzahl	8997	8302	4241	4277
Durchschnittsalter	10.5	10.5	10.4	10.3
Mädchen (%)	50	49	49	49
Familie mit hohem sozioökonomischen Status (%) (mindestens ein Elternteil ist Akademiker oder Führungskraft)	29	29	32	29
Kinder mit Migrationshintergrund (%) (beide Elternteile im Ausland geboren)	14*	15*	16	19
Kinder mit besonderen Unterstützungsbedarfen (%)	-	-	5*	7
Testteilnahme				
Anzahl	7633	7899	4000	3959
Teilnahmequote (%)	88	94	96	95
Beteiligung am Schülerfragebogen (%)	98	96	91	88
Eltern				
Beteiligung am Elternfragebogen (%)	89	87	80	72

* = Unterschied zu IGLU 2016 statistisch signifikant ($p < .05$).

IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

8 Aufbereitung und Analyse der Daten

In Deutschland war, wie auch in den vergangenen Erhebungsrunden, die *IEA Hamburg* mit der Organisation der nationalen Datenverarbeitung, -aufbereitung und der Aufgabenkodierung beauftragt. Die Dateneingabe und -verarbeitung erfolgte technologiegestützt, so dass die Testhefte elektronisch eingelesen werden konnten. Neben einer niedrigen Fehlerquote im Prozess der Datenverarbeitung hat dies den Vorteil, auch zu einem späteren Zeitpunkt schnell und problemlos auf die Antworten einzelner Schülerinnen und Schüler zugreifen zu können, beispielsweise wenn individuelle Antworten überprüft werden müssen. Antworten, die von der Scan-Software nicht sicher erkannt werden konnten, wurden von ausgebildeten Mitarbeiterinnen und Mitarbeitern nachträglich verifiziert. Die Datensätze wurden abschließend mit Hilfe einer speziell entwickelten Software kritisch auf ungültige Daten und Inkonsistenzen geprüft.

Kodierung der Leistungstests

Die Schülerlösungen der bearbeiteten offenen Testaufgaben wurden nach eng umrissenen internationalen Vorgaben bewertet und für die Analysen kodiert. Während bei Mehrfachwahlaufgaben die richtige Lösung eindeutig festgelegt ist, weisen Testfragen mit offenem Antwortformat (siehe Abschnitt 5.1.1 sowie Kapitel 3) einen kleinen Bewertungsspielraum auf.

In Deutschland wurde die Kodierung der Fragen mit offenem Antwortformat der *IEA Hamburg* übertragen und dort von erfahrenen Kodiererinnen und Kodierern übernommen, bevorzugt von solchen, die Deutsch studieren. Zur Vorbereitung wurden die Kodiererinnen und Kodierer mit Beispiellösungen intensiv für ihre Aufgabe geschult. Sie erhielten eine Kodieranweisung mit genauen Beschreibungen für richtige und falsche Lösungen zu jeder einzelnen Aufgabe. Die Güte der Kodierungen wurde anhand von Doppelkodierungen geprüft. Dazu wurden 200 zufällig ausgewählte Antworten zu jeder Aufgabe von einer zweiten Person beurteilt und es wurde die Übereinstimmung zwischen der Erst- und der Zweitkodierung bestimmt. In IGLU 2016 ist, wie auch in den Erhebungszyklen zuvor, die durchschnittliche Übereinstimmung über alle Aufgaben hinweg in allen Staaten sehr hoch (Martin et al., 2017).

9 Skalierung der Leistungstests

Mit der in IGLU eingesetzten Leistungstestung soll von den in den Testheften gezeigten Schülerantworten auf die zu messende Kompetenz von Viertklässlerinnen und Viertklässlern geschlossen werden. Die Verortung von Schülerinnen und Schülern auf einer Kompetenzskala wird Skalierung genannt. Zur Bestimmung individueller Leistungswerte auf den Kompetenzdomänen wird in IGLU auf statistische Modelle zurückgegriffen, die auf der *Item Response Theory* basieren (IRT, Boomsma, van Duijn & Snijders, 2000; Linden v. d. & Hambleton, 1997). Die IRT geht von der Annahme aus, dass die beobachtbaren Antworten einer Person in einem Test durch eine nicht beobachtbare, testbezogene Fähigkeit erklärbar sind. Die beobachtbaren Antworten werden verwendet, um die Höhe der nicht beobachtbaren (*latenten*) Fähigkeiten zu schätzen. Mit der IRT können die Schwierigkeiten der Aufgaben und die Kompetenzen der Personen (auch als *Fähigkeiten* bezeichnet) auf derselben Skala abgebildet werden, das heißt die Schwierigkeiten der Aufgaben und die Fähigkeiten der Personen sind auf einer Metrik direkt nebeneinander vergleichbar. Damit ist zudem möglich, Personen die unterschiedliche Aufgaben derselben Metrik bearbeitet haben, hinsichtlich ihrer Fähigkeiten zu vergleichen.

Die beschriebene Eigenschaft von IRT-Modellen macht dieses Verfahren für IGLU (und vergleichbare Studien wie TIMSS) besonders interessant. Bedingt durch das in IGLU implementierte Testheftdesign (Multi-Matrix-Design, siehe Tabelle 2.3) bearbeiten einzelne teilnehmende Schülerinnen und Schüler immer nur eine Auswahl der Testaufgaben des IGLU-Aufgabenpools (Foy, Brossman & Galia, 2011). Durch die Etablierung einer IRT-Skala für den IGLU-Aufgabenpool wird gewährleistet, dass dennoch ein direkter Vergleich dieser Schülerleistungen möglich ist – eine ausreichend große Stichprobe und Überlappungen der Testheftinhalte ist dabei vorausgesetzt (z. B. Kolen, 1981; Kolen & Brennan, 2004; Lord, 1980). IGLU zielt nicht auf Individualdiagnostik, sondern dient ausschließlich der Beschreibung von Kompetenzverteilungen in den untersuchten Populationen. Da diese Verteilungen möglichst genau geschätzt werden sollen, wird in IGLU

zur Ermittlung der Schülerleistung (*Personenparameter*) der *Plausible-Value*-Ansatz gewählt (Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992); ein Verfahren, dem die Theorie der Multiplen Imputation zugrunde liegt (Rubin, 1987).

Die Grundidee des *Plausible-Value*-Ansatzes ist, die nicht beobachtbare wahre Fähigkeit einer Person als prinzipiell unbekannt zu betrachten. Diese nicht bekannte Fähigkeit wird dann durch ‚plausible Werte' abgebildet. Zur Bestimmung der plausiblen Werte wird neben der Information über die Testleistung der Person eine Vielzahl von Informationen aus den Hintergrundfragebögen mitberücksichtigt. Auf Basis dieser Information wird eine (bedingte) Wahrscheinlichkeitsverteilung für die Fähigkeit einer Person erstellt. *Plausible Values* sind jene Werte, die aus dieser bedingten Verteilung per Zufall gezogen werden. Durch die Ziehung mehrerer *Plausible Values* wird berücksichtigt, dass die Bestimmung eines fehlenden Wertes immer auch mit Unsicherheit behaftet ist. Als Konvention hat sich die Ziehung von fünf Werten etabliert. Zur angemessenen Schätzung von Populationskennwerten sind alle *Plausible Values* zu berücksichtigen, und bei der Bestimmung der Stichprobenunsicherheiten auch deren Varianz. Analysen mit den Leistungsdaten müssen entsprechend immer fünfmal erfolgen und die Ergebnisse dann zusammengefasst werden.

Skalierungsmodelle

Im vorliegenden Bericht werden die Kompetenzen der Schülerinnen und Schüler auf Grundlage der Skalierung berichtet, die die internationale Studienleitung durchgeführt hat (Martin et al., 2017). Die Skalierung der Leistungsdaten aller Teilnehmerstaaten ist komplex und zeitintensiv, ihr gehen umfangreiche Datenprüfungs- und Aufbereitungsschritte voraus. Für IGLU 2016 erfolgte erst nach Abschluss aller Prüfungen und einem Review der Ergebnisse durch alle beteiligten nationalen Studienleitungen die Weitergabe der Datensätze aller Teilnehmerstaaten und Regionen an die nationalen Studienleitungen im August 2017. Zentrale Modellparameter werden in der technischen Dokumentation zur Studie veröffentlicht (ebd.).

Die internationale Skalierung erfolgte im Zeitraum von Herbst 2016 bis Sommer 2017. Die Qualitätssicherung der Berechnungen wurde durch Expertinnen und Experten des US-amerikanischen *Educational Testing Service* (ETS) durchgeführt. Ausgangspunkt der Skalierung ist die umfassende Prüfung der Messeigenschaften jeder einzelnen Testaufgabe nach festgelegten Kriterien (ebd.). Im Ergebnis dieser Prüfungen standen für die Ermittlung der Kompetenzmittelwerte im Lesen 169 administrierte Testaufgaben zur Verfügung.

Die Bestimmung der *Plausible Values* erfolgte zunächst unter Nutzung eines mehrdimensionalen dreiparametrigen logistischen Modells (3-PL-Modell), in dem neben der Schwierigkeit der Aufgabe auch deren Trennschärfe und (bei Aufgaben im *Multiple-Choice*-Format) ein Parameter für die Ratewahrscheinlichkeit berücksichtigt wird. Geschätzt wurden die Modellparameter mit dem Computerprogramm *Parscale* (Muraki & Bock, 1999) unter Verwendung des *Marginal-Maximum-Likelihood* (MML) Ansatzes. Anschließend wurde für die Schätzung der Personenparameter (*Plausible Values*) die Software *MGROUP* (Sheehan, 1985) verwendet. Um genauere Schätzer der Personenparameter zu erhalten, kommt in IGLU 2016 für die Skalierung der Subskalen (wie auch schon in IGLU 2011) ein mehrdimensionales Antwortmodell zur Anwendung, mit dem die in den Daten enthaltene Zusammenhangsstruktur besser abgebildet werden kann. Zuvor, das heißt in den vorangegangenen Erhebungszyklen 2001 und 2006, wur-

den die Kompetenzwerte für die Subskalen (je zwei Leseintentionen und zwei Verstehensprozesse, siehe Kapitel 3 in diesem Band) separat ermittelt. Um die Vergleichbarkeit zu den vorangegangenen Studienzyklen sicherzustellen wurden im vorliegenden Berichtsband die im Rahmen der nationalen Berichtslegung von IGLU 2011 neu skalierten Daten zu IGLU 2001 und 2006 genutzt (vgl. auch Tarelli et al., 2012). Abweichungen zur Berichtslegung zu IGLU 2001 und 2006 (Bos et al., 2003; Bos et al., 2007; Mullis, Martin, Gonzalez & Kennedy, 2003; Mullis et al., 2007) sind demnach durch die Veränderung des Skalierungsmodells bedingt.

10 Gewichtung und Schätzung von Stichproben- und Messfehlern

Das wesentliche Merkmal einer repräsentativen Stichprobe besteht darin, von den Verhältnissen in einer Stichprobe auf die Grundgesamtheit schließen zu können. Dies setzt voraus, dass die Stichprobe nicht verzerrt und die Grundgesamtheit angemessen repräsentiert ist. In IGLU finden sich zwei Ursachen, die zu Stichprobenverzerrungen führen können: Erstens hat aufgrund des Designs der Stichprobenauswahl (siehe Abschnitt 7) nicht jede Schülerin beziehungsweise jeder Schüler dieselbe Wahrscheinlichkeit, in die Stichprobe zu gelangen. So hängt die Wahrscheinlichkeit unter anderem von der Zügigkeit der Schulen ab: In einer Schule mit zwei Jahrgangsklassen hätte ein Schulkind eine Wahrscheinlichkeit von 50 Prozent in die Stichprobe zu gelangen. In einer Schule mit vier Jahrgangsklassen wäre es dagegen nur eine Wahrscheinlichkeit von 25 Prozent. Zweitens kann auch der Stichprobenausfall eine Verzerrung bedeuten. Nur selten sind Ausfälle zufällig; vielmehr weisen sie oft einen Zusammenhang mit studienrelevanten Merkmalen auf. So ist es denkbar, dass leistungsschwache Schülerinnen und Schüler mit höherer Wahrscheinlichkeit am Testtag fehlen als durchschnittliche oder leistungsstarke Schülerinnen und Schüler. Werden Stichprobenverzerrungen nicht angemessen korrigiert, können inferenzstatistische Methoden zu falschen Schlussfolgerungen führen.

In IGLU wird diesem Problem dadurch begegnet, dass für jedes getestete Schulkind ein statistisches Gewicht berechnet wird. Die Gewichte werden dann bei der Berechnung aller Statistiken verwendet, zum Beispiel bei einem Mittelwert oder bei Prozentangaben. Die Daten der teilnehmenden Schülerinnen und Schüler tragen dadurch in kontrollierter Art und Weise unterschiedlich stark zu der Berechnung der Statistiken bei.

Da in IGLU keine reinen Zufallsstichproben vorliegen, sondern sogenannte Clusterstichproben gezogen werden, kann die Bestimmung des Standardfehlers nicht mit sonst üblichen Analyseverfahren vorgenommen werden, da sonst der Standardfehler systematisch unterschätzt werden würde. Eine präzisere Bestimmung des Standardfehlers erlauben sogenannte *Jackknife*-Verfahren. Diese Verfahren bestimmen die Variabilität der Schätzung von Populationskennwerten (wie z. B. die Leseleistung der in IGLU 2016 getesteten Grundschülerinnen und Grundschüler) durch ein wiederholtes Schätzen dieser Werte aus Substichproben, was die Möglichkeit bietet, Stichprobenfehler zu schätzen, ohne zugleich die Annahme einfacher Zufallsstichproben voraussetzen zu müssen. Durch die Reduzierung der Freiheitsgrade führt diese Methode zu akkurateren Schätzungen der Standardfehler.

Die korrekte Bestimmung des Standardfehlers ist sehr wichtig, weil er genutzt wird, um zu ermitteln, ob sich zwei Gruppen signifikant voneinander unterscheiden. Im vorliegenden Bericht werden entsprechend alle Standardfehler mit solch einem Verfahren (*Jackknife Repeated Replication Technique,* JRR) geschätzt. In diesem Bericht wird für die Entscheidung über die Signifikanz eine Irrtumswahrscheinlichkeit von $\alpha = .05$ zugrunde gelegt.

11 Zur Darstellung und Interpretation der Ergebnisse

Zur Darstellung der im vorliegenden Band berichteten Ergebnisse in Tabellen und Abbildungen werden verschiedene statistische Kennwerte verwendet. Die wichtigsten Kennwerte sind nachfolgend erläutert. Darüber hinausgehende Begriffserklärungen und technische Grundlagen sind in dem technischen Bericht der internationalen Studienleitung (Martin et al., 2017) sowie in einschlägiger Fachliteratur dokumentiert.

Mittelwerte und Standardabweichungen von Leistungsdaten
Die in diesem Bericht dokumentierten Ergebnisse und Vergleiche basieren auf der internationalen Stichprobe und den internationalen Kompetenz- und Fragebogenskalen von IGLU 2016 (Martin et al., 2017; Mullis & Martin, 2015). Zur Darstellung der Leistungswerte wurde für die erste Erhebung von IGLU 2001 und für alle nachfolgenden Zyklen von IGLU ein Mittelwert (*M* für arithmetisches Mittel) von 500 Punkten und eine Standardabweichung (*SD* für *Standard Deviation*) von 100 Punkten festgelegt (Mullis et al., 2003; Martin et al., 2003). Die Wahl der Einheiten für diese Skala basiert auf Konventionen. Werte, die nahe beim Mittelwert liegen, kommen häufiger vor als Extremwerte. Oft ergibt sich eine Normalverteilung, wie sie in Abbildung 2.5 (siehe Seite 68) dargestellt ist.

Die durchschnittliche Streuung der Werte um den Mittelwert wird durch die Standardabweichung statistisch gekennzeichnet. Im Bereich einer Standardabweichung über und unter dem Mittelwert (d. h. in Abbildung 2.5 im Bereich von 400 bis 600 Punkten) liegen rund zwei Drittel (68.3 %) aller Testwerte der internationalen Population. Bei zwei Standardabweichungen erhöht sich dies auf 95.5 Prozent und bei drei Standardabweichungen auf 99.7 Prozent.

Differenzen und damit Veränderungen von Leistungsmittelwerten zwischen den Erhebungszyklen, wie sie im Trend berichtet werden, werden im vorliegenden Band mitunter durch das Symbol Delta (Δ) gekennzeichnet.

Internationaler Mittelwert versus Skalenmittelwert
IGLU ist als Trendstudie konzipiert, um Veränderungen in den mittleren Leistungsniveaus von Schülerinnen und Schülern der teilnehmenden Staaten und Regionen über die Zeit hinweg darstellen zu können. Der Darstellung der Ergebnisse liegt dieselbe Skala zugrunde, damit die Schülerleistungen, die zu den verschiedenen Zyklen erzielt wurden, miteinander verglichen werden können. Die Vergleichbarkeit der Ergebnisse wird sichergestellt, indem in jedem Zyklus Aufgaben aus den vorangegangenen Studienzyklen erneut eingesetzt werden (siehe Abschnitt 5.1). Die in IGLU 2016 ermittelten Ergebnisse und Daten zu Testaufgaben, die zu mehreren Erhebungszyklen eingesetzt wurden, können in einer gemeinsamen Skalierung mit den Daten von 2011 verankert werden (Martin et al., 2017; Foy et al., 2011). In gleicher Weise lassen sich die Daten

aus IGLU 2011 wiederum zu denen aus IGLU 2006 beziehungsweise IGLU 2001 in Beziehung setzen. Letztlich bilden die Daten der ersten IGLU-Erhebung im Jahr 2001 den Referenzpunkt für alle Folgezyklen. Veränderungen in den Leistungen zwischen den einzelnen Zyklen können auch 2016 auf der Basis des in IGLU 2001 bestimmten Mittelwerts von $M = 500$ verglichen und beurteilt werden. Dieser Referenzwert wird in IGLU als *Skalenmittelwert* bezeichnet.

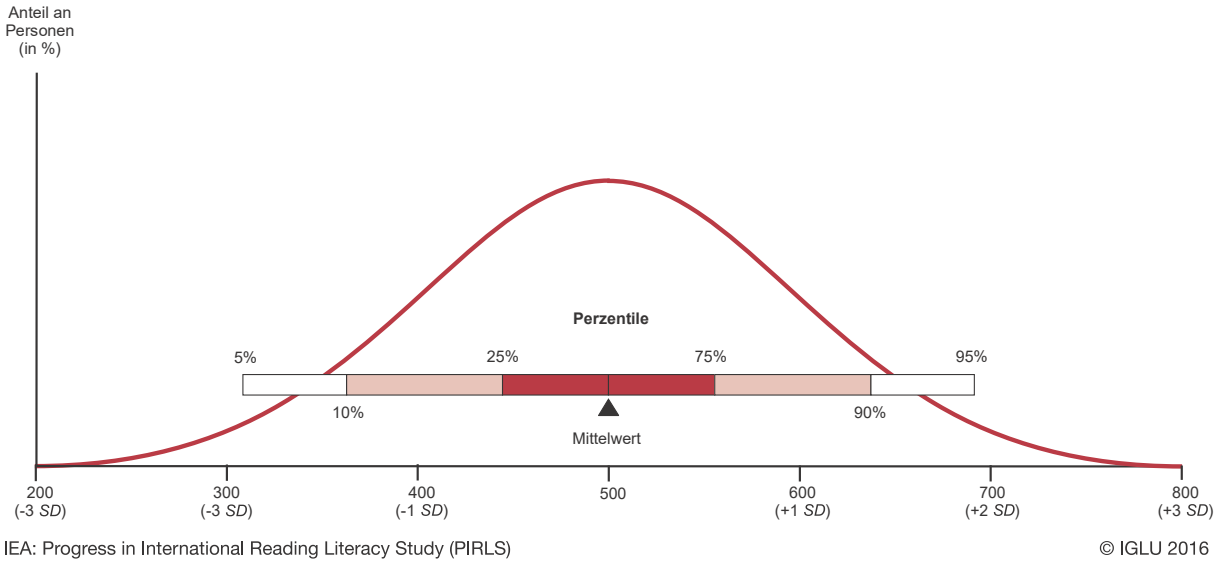
Der *internationale Mittelwert* bezeichnet den Wert, der mit jeder IGLU-Erhebung neu berechnet wird: Es handelt sich dabei um den Mittelwert, der aus allen Mittelwerten der jeweiligen Teilnehmerstaaten eines Zyklus gebildet wird (ohne Berücksichtigung der Benchmark-Teilnehmer, siehe Abschnitt 4.2.1). Im Gegensatz zum Skalenmittelwert variiert der internationale Mittelwert zwischen den Erhebungszyklen. Setzen sich zum Beispiel die Teilnehmer einer Erhebung aus Staaten zusammen, die im Vergleich zu jenen Staaten, die an IGLU 2001 teilgenommen haben, insgesamt leistungsstärker sind, ergibt sich ein internationaler Mittelwert der größer als 500 ist. In IGLU 2016 beträgt der internationale Mittelwert für Lesen 521 Punkte ($SD = 78$). Die an IGLU 2016 teilnehmenden Staaten sind im Durchschnitt leistungsstärker als die Teilnehmer an IGLU im Jahr 2001. Bei der Berichterstattung zu den Hintergrunddaten wird nicht zwischen einem internationalen Mittelwert und einem Skalenmittelwert unterschieden, weil die Fragebogeninhalte im Gegensatz zu den Leistungstests in sich heterogene Inhalte abdecken, die von Erhebung zu Erhebung teilweise variieren.

Median, Perzentile und Perzentilbänder

Neben dem Mittelwert und der Standardabweichung werden auch Perzentilwerte berichtet. Perzentile informieren, ebenso wie die Standardabweichung, über die Variation der Werte. Ein Perzentilwert gibt an, wie viel Prozent der untersuchten Personen einen Wert erreichen oder darunter bleiben. Liegt beispielsweise der Leistungswert zum 5. Perzentil bei 318, bedeutet dies, dass 5 Prozent der untersuchten Schülerinnen und Schüler einen Punktwert von 318 oder geringer erreichen. Zugleich bedeutet dies aber auch, dass 95 Prozent der Schülerinnen und Schüler einen Wert erreichen, der höher als 318 ist. Entsprechend trennt der Punktwert des 25. Perzentils das untere Leistungsviertel ab, der des 75. Perzentils das obere Leistungsviertel. Das 50. Perzentil, auch Median genannt, trennt die Verteilung in zwei Hälften mit je gleicher Personenzahl.

Im vorliegenden Bericht werden die Perzentilwerte tabellarisch oder graphisch in Form von Perzentilbändern (siehe Abbildung 2.5) dargestellt. Passen sich die Werte einer Normalverteilung an (siehe ebd.), ergibt sich ein symmetrisches Perzentilband. In diesem Fall fällt auch der Median mit dem Mittelwert zusammen.

Abbildung 2.5: Normalverteilung mit Perzentilen



Kompetenzstufen

Die Leistungsskala in IGLU deckt ein breites Kompetenzspektrum ab. Die in IGLU eingesetzten Lesetexte und Testaufgaben thematisieren verschiedene Inhalte und aktivieren verschiedene Prozesse kognitiver Anforderungen (siehe Abschnitt 4.1.1). Um die erreichten Kompetenzwerte der getesteten Schülerinnen und Schüler inhaltlich interpretieren zu können, werden in IGLU Kompetenzstufen gebildet. Ausführlich beschrieben und anhand von Beispielaufgaben illustriert sind die Kompetenzstufen in Kapitel 3 in diesem Band.

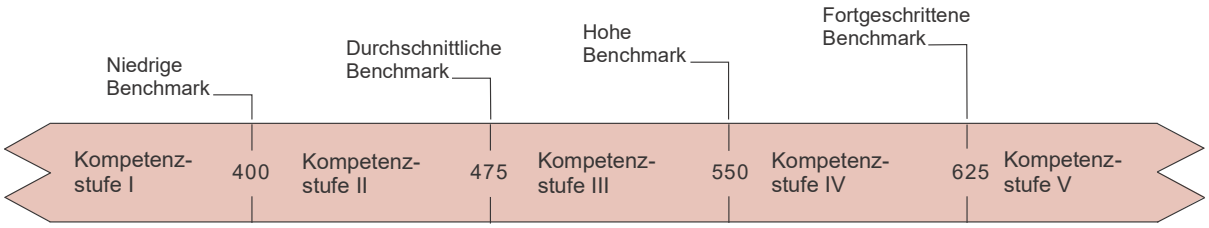
Die Entwicklung und Beschreibung der Kompetenzstufen wird von einem international zusammengesetzten Expertengremium unter der Leitung der internationalen Studienleitung vorgenommen. Bei der Entwicklung von Kompetenzstufen wurden als zentrale Bezugspunkte auf den Leistungsskalen sogenannte *Benchmarks* festgelegt (Mullis & Martin, 2015; Martin, Mullis, Foy & Stanco, 2012). Die vier *Benchmarks* (400, 475, 550, 625) teilen die Leistungsskala in fünf Abschnitte, die in Deutschland als *Kompetenzstufen* bezeichnet werden. Die niedrige Benchmark liegt bei 400 Punkten. Sie beschreibt den Grenzbereich zwischen Kompetenzstufe I und Kompetenzstufe II. Die durchschnittliche *Benchmark* liegt bei 475 Punkten und markiert den Beginn von Kompetenzstufe III. Die hohe und die fortgeschrittene *Benchmark* liegen bei 550 beziehungsweise 625 Punkten. Mit ihrer Überschreitung beginnen Kompetenzstufe IV beziehungsweise Kompetenzstufe V (siehe Abbildung 2.6).

Die Anordnung der internationalen *Benchmarks* und die sich daraus ergebenden Intervalle für die Kompetenzstufen sind in Abbildung 2.6 grafisch dargestellt. Das unter der niedrigen *Benchmark* liegende Intervall umfasst in Deutschland Leistungen der Kompetenzstufe I. Schülerinnen und Schüler, die Kompetenzstufe I erreichen, sind nicht in der Lage, die für Kompetenzstufe II charakteristischen, dieser Schülerkohorte entsprechend relativ einfachen Aufgaben zu lösen. Das Kompetenzniveau von Schülerinnen und Schülern am unteren Ende der Leistungsskala lässt sich auf der Basis der eingesetzten IGLU-Leistungstests und aufgrund der geringen Anzahl richtig gelöster Aufgaben nicht differenziert beschreiben. Interpretationen zur Schülergruppe auf diesem Kompetenzniveau sind daher zunächst in dem Sinne möglich, dass lediglich aus-

gesagt werden kann, dass Schülerinnen und Schüler dieser Kompetenzstufe nicht die mit dem internationalen IGLU-Test festgelegten Mindestanforderungen erreichen. Die Erweiterung um ELFE II (Untertests auf Wort- und Satzebene) in IGLU 2016 ermöglicht erstmals, die Leseleistungen dieser Schülerinnen und Schüler auf Wort- und Satzebene konkreter zu beschreiben (siehe Abschnitt 5.1.4).

Der Beschreibung und Entwicklung der Kompetenzstufen liegt eine spezifische Auswahl von Aufgaben zugrunde, die für das auf den vier Benchmarks gezeigte Leistungsniveau charakteristisch sind. Um zu dieser Aufgabenauswahl zu gelangen, werden Schülerinnen und Schüler ausgewählt, die folgende mittlere Leistungswerte erreichten: 395–405, 470–480, 545–555, 620–630. Die Leistungen dieser Schülerinnen und Schüler lassen sich damit im Bereich von fünf Punkten unter bis fünf Punkte über einem *Benchmark*-Wert verorten. Diese Zuordnung ist aufgrund eines Verfahrens zulässig, das die Darstellung von Schülerfähigkeiten und Aufgabenschwierigkeiten auf einer gemeinsamen Skala erlaubt (siehe Abschnitt 9). Zur Beschreibung mittlerer *Benchmark*s werden Aufgaben herangezogen, die von mindestens 65 Prozent der zugeordneten Schülerinnen und Schüler gelöst werden, zugleich aber von weniger als 50 Prozent der Schülerinnen und Schüler der darunter liegenden *Benchmark*. Für die Kompetenzstufen am oberen und unteren Ende der Leistungsskala gelten diesem Vorgehen folgend leicht angepasste Kriterien (Mullis & Martin, 2015). In Tabellen und Abbildungen dieses Bandes wird für Kompetenzstufen die Abkürzung ‚KS' genutzt, wo dies der besseren Darstellbarkeit dient.

Abbildung 2.6: Beziehung von Benchmarks und Kompetenzstufen



IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

Standardfehler und Signifikanz

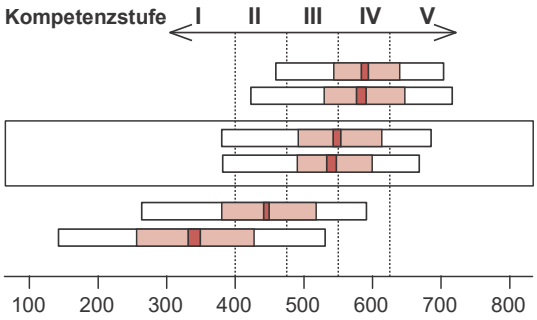
Die Repräsentativität einer Stichprobe, wie sie in IGLU gegeben ist, stellt sicher, dass die anhand dieser Stichprobe erlangten Erkenntnisse auf die Grundgesamtheit übertragen werden können. Es gibt jedoch in jeder Stichprobe kleinere oder größere Abweichungen von der Grundgesamtheit. In der Regel ist die Abweichung umso geringer, je größer die gewonnenen Stichproben ausfallen. Berechnet man beispielsweise den Mittelwert für jede dieser Stichproben, lässt sich für diese Mittelwerte ähnlich wie für die Messwerte innerhalb einer Stichprobe eine Streuung berechnen. Diese Streuung beziehungsweise das berechnete Streuungsmaß bezeichnet den Standardfehler (*SE* für *Standard Error*) oder auch Standardschätzfehler.

Der Standardfehler quantifiziert die Unsicherheit der aus Stichprobendaten geschätzten Populationswerte. Je geringer er ausfällt, desto genauer ist der geschätzte Populationskennwert. Für den Standardfehler gilt, dass im Bereich von rund zwei Standardfehlern (1.96) unter beziehungsweise über einem errechneten

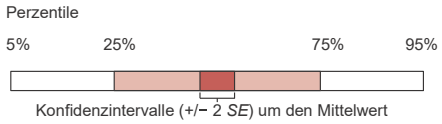
Mittelwert mit 95-prozentiger Wahrscheinlichkeit der ‚wahre' Populationswert liegt.

Für die Interpretation der Ergebnisse stellt der Standardfehler ein wichtiges Maß dar. Er dient der Einschätzung, ob sich zwei oder mehr Gruppen signifikant voneinander unterscheiden. Im vorliegenden Bericht werden (wie auch in den vergangenen Berichten zu IGLU 2001, 2006 und 2011) Perzentilbänder zur Ergebnisdarstellung angegeben, denen Konfidenzintervalle der Mittelwerte und dabei berechnete Standardfehler zugrunde liegen. Die Konfidenzintervalle (Vertrauensintervalle) der Mittelwerte (siehe Abbildung 2.7) geben an, in welchem Wertebereich der wahre Populationswert mit 95-prozentiger Wahrscheinlichkeit liegt. Zugrunde gelegt wird das Intervall von über beziehungsweise unter 1.96 Standardfehlern um die jeweilige Statistik. Überlappen sich beispielsweise zwei Konfidenzintervalle nicht, wie in Abbildung 2.7 die Konfidenzintervalle um die mittlere Leseleistung der Russischen Föderation oder Singapur und Deutschland, entspricht dies einem signifikanten Unterschied mit einer Irrtumswahrscheinlichkeit von $\alpha = .05$.

Abbildung 2.7: Darstellung von Perzentilbändern mit Konfidenzintervallen am Beispiel der Leseleistung in IGLU 2016



	Teilnehmer	*M*	(*SE*)	*SD*	(*SE*)
	Russische Föderation	581	(2.2)	66	(1.3)
2	*Singapur*	576	(3.2)	80	(2.1)
	Australien	544	(2.5)	84	(1.6)
	Deutschland	537	(3.2)	78	(3.2)
1 2	*Malta*	452	(1.8)	90	(1.5)
7	*Marokko*	358	(3.9)	107	(1.7)



☐ Nicht statistisch signifikant vom deutschen Mittelwert abweichende Staaten ($p > .05$).
Kursiv gesetzt sind die Teilnehmer, für die von einer eingeschränkten Vergleichbarkeit der Ergebnisse ausgegangen werden muss.
1 = Die nationale Zielpopulation entspricht nicht oder nicht ausschließlich der vierten Jahrgangsstufe.
2 = Der Ausschöpfungsgrad und/oder die Ausschlüsse von der nationalen Zielpopulation erfüllen nicht die internationalen Vorgaben.
7 = Teilnahme an IGLU 2016 und PIRLS Literacy (Iran, Marokko) bzw. ausschließlich an PIRLS Literacy (Dänemark, 3. Jgst.).
Die Kennwerte für Iran und Marokko werden in Anlehnung an die internationale Berichterstattung als Mittelwert der beiden Studien dargestellt.

IEA: Progress in International Reading Literacy Study (PIRLS) © IGLU 2016

Mehrebenenanalysen

Die Betrachtung der Leseleistung im Trend zeigt für die einzelnen teilnehmenden Staaten und Regionen unterschiedliche Verläufe; bei manchen Staaten und Regionen lassen sich positive Veränderungen nachzeichnen, bei anderen keine oder gar negative (siehe Kapitel 3 in diesem Band). Über die Zeit haben sich neben der Testleistung aber auch die Rahmenbedingungen (z. B. Anteil der Schülerinnen und Schülern mit Migrationshintergrund in der Population) geändert. Eine Möglichkeit, die Bedeutung dieser Veränderungen in den Rahmenbedingungen für die Leistungsentwicklungen von Viertklässlerinnen und Viert-

klässlern im Lesen abzuschätzen und damit quasi einen Nettoeffekt der Trendentwicklung zu ermitteln, stellt die Anwendung von Trendmodellen dar. Ermittelt wird die Leistungsdifferenz zwischen den Studienzyklen, die sich ergeben hätte, wenn die Verteilung der soziodemografischen Merkmale in der Schülerschaft gleichgeblieben wäre. Dieses Verfahren wurde beispielsweise von der OECD in ihrem internationalen Bericht über Trends in der Lesekompetenz zwischen 2000 und 2009, die im Rahmen des *Programme for International Student Assessment* (PISA) ermittelt wurden, angewendet (OECD, 2010, S. 49 ff.; Ehmke, Klieme & Stanat, 2013). Um die Leistungsveränderungen über die Zeit nach Kontrolle von Merkmalen zu berücksichtigen, wurde pro teilnehmendes Land eine lineare Regression für die Schülerleistungen bestimmt. Als erklärende Faktoren der Leistung wurden der Messzeitpunkt, die soziodemographischen Merkmale und die Interaktionen zwischen den soziodemographischen Merkmalen und dem Messzeitpunkt berücksichtigt. In einem ähnlichen Vorgehen wurden auch im Rahmen der nationalen Berichtslegung zu TIMSS 2015 Trendanalysen für Deutschland vorgestellt (vgl. Wendt et al., 2016; darin insbesondere Kasper, Wendt, Bos & Köller, 2016).

In diesem Berichtsband wird zur Betrachtung der Leistungsentwicklung im Trend in Deutschland ein methodisches Vorgehen gewählt, das demjenigen der OECD-Studie und demjenigen in TIMSS 2015 in wesentlichen Teilen entspricht. Wie schon für die Trendanalysen in TIMSS 2015 für Deutschland wurde in Abweichung zu dem Vorgehen in der OECD-Studie kein lineares Regressionsmodell verwendet, sondern ein lineares Modell mit Zufallseffekten (ein sogenanntes Mehrebenenmodell; McCulloch, Searle & Neuhaus, 2008). Die Verwendung dieses Modells anstelle der linearen Regression scheint in dem vorliegenden Kontext, wie auch in TIMSS 2015, aus methodischer Sicht geboten. Eine Anwendung der linearen Regression würde voraussetzen, dass die Leistungen der Schülerinnen und Schüler nach Kontrolle aller erklärenden Variablen unabhängig voneinander sind (Werner, 1997). Allerdings ist aus der empirischen Bildungsforschung bekannt, dass die individuelle Schülerleistung auch vom Klassen- beziehungsweise Schulkontext abhängt. Die Leistungen von Schülerinnen und Schülern innerhalb einer Klasse und Schule sind also im Allgemeinen nicht unabhängig voneinander. Da die Schülerinnen und Schüler in der IGLU-Stichprobe im Klassenverbund getestet werden (siehe Abschnitt 4.3), ist davon auszugehen, dass auch ihre Leistungen in der IGLU-Stichprobe (ohne Kontrolle der Klassen- und Schulzugehörigkeit) voneinander abhängig sind. Das verwendete Mehrebenenmodell berücksichtigt diese Abhängigkeiten in den Schülerleistungen: Indem die klassenbedingten Gemeinsamkeiten in der Leistung durch einen Interceptwert, der allen Schülerinnen und Schülern einer Klasse gemeinsam ist, modelliert und repräsentiert werden, können die Abhängigkeiten zwischen den Schülerleistungen kontrolliert werden. Da die untersuchten Klassen außerdem eine Zufallsauswahl aus allen zur definierten Zielpopulation gehörenden Klassen in Deutschland darstellen, konnte die konkrete Ausprägung des Interceptwerts als zufällig betrachtet werden, weswegen ein sogenanntes *random intercept model* berechnet wurde (Raudenbush & Bryk, 2002).

Anders als in der OECD-Studie, aber vergleichbar zum Vorgehen in TIMSS 2015, basieren die Mehrebenenmodelle auf nicht imputierten Datensätzen. Dieses Vorgehen ermöglicht es, dass die Ergebnisse der hier berichteten Analysen mit den Trendergebnissen im internationalen Vergleich aus den Kapiteln in diesem Band abgeglichen werden können: Abweichungen zwischen den Trendschätzern in diesem und den anderen Kapiteln aufgrund eines unterschiedlichen Umgangs

mit fehlenden Werten sind demnach auszuschließen. In Bezug auf den Studienzyklus 2011 wurde der kombinierte Datensatz aus TIMSS und IGLU verwendet, um eine Vergleichbarkeit zu den Analysen der im Rahmen der nationalen Berichterstattungen publizierten Ergebnisse zu den Hintergrundmerkmalen der Schülerinnen und Schüler gewährleisten zu können (vgl. Bos, Tarelli et al., 2012; Bos, Wendt et al., 2012).

Das hier beschriebene Verfahren entspricht weitestgehend dem methodischen Vorgehen von Van Damme und Bellens (2016), die die Trendentwicklung in TIMSS für ausgewählte Länder analysiert haben. Die Mehrebenenanalysen wurden mit dem Softwareprogramm SAS/STAT Software, Version 9.4 (TS1M1) von SAS System für Windows durchgeführt. Durch die Implementierung einer eigens für diese Analysen entwickelten Funktion in dem Programm konnten die *Sampling*-Varianz und die Varianz zu Lasten der *Plausible Values* bei den Analysen berücksichtigt werden.

Staatenvergleiche

Durch die weltweite Beteiligung von 57 Bildungssystemen an IGLU 2016 (siehe Abschnitt 4.2.1), eröffnen sich zahlreiche Möglichkeiten, die Leistungen der Schülerinnen und Schüler in Deutschland mit denen von Kindern in anderen Staaten und Regionen zu vergleichen. Zu klären ist dabei immer die Relevanz solcher Vergleiche sowie die Frage, in welcher Weise Vergleiche sinnvoll sind oder ob sie aufschlussreiche Informationen liefern. So ist etwa zu fragen, in welchem Ausmaß ein Vergleich von Bildungssystemen überhaupt tragfähig ist. Einen Überblick zu dieser Thematik bietet die von der internationalen Studienleitung veröffentlichte *Enzyklopädie* (siehe Abschnitt 4.1.2), in der nachzulesen ist, wie die Bildungssysteme der Teilnehmer strukturiert und inhaltlich ausgestaltet sind (Mullis et al., 2017). Die Beschreibung der Bildungssysteme folgt einer einheitlich vorgegebenen Gliederung, so dass relevante Informationen systematisch strukturiert für alle Teilnehmer vorliegen (für Deutschland siehe Wendt, Walzebug, Bos, Smith und Bremerich-Vos, 2017). Neben inhaltlichen Fragen ist des Weiteren zu klären, ob unter Berücksichtigung nationaler Besonderheiten der Stichproben (siehe Abschnitt 4.3) der Vergleich einzelner Staaten zulässig ist. So ist ein Vergleich mit Regionen wie den Benchmark-Teilnehmern nur unter Einschränkungen möglich (siehe Abschnitt 4.2.1), da bei Benchmark-Teilnehmern Besonderheiten eines Schulwesens im Vordergrund stehen könnten, die möglicherweise nicht für einen ganzen Staat zutreffen. Die internationale Studienleitung weist deshalb in Absprache mit den nationalen Studienleitungen Benchmark-Teilnehmer stets getrennt in den Ergebnisdarstellungen aus.

Um sinnvolle Vergleiche zu Deutschland in der vorliegenden Berichterstattung dokumentieren zu können, werden in IGLU Vergleichsgruppen gebildet (siehe Abschnitt 4.2.1, Tabelle 2.1). Ihnen gehören Staaten oder Staatengruppen an, die sich hinsichtlich des kulturellen und ökonomischen Hintergrunds oder der wirtschaftlichen Situation ähneln und sich daher für einen Staatenvergleich mit Deutschland eignen. Aus bewährten Gründen (vgl. Bos et al., 2003; Bos et al., 2007; Bos, Tarelli et al., 2012) stehen auch in diesem Bericht zu IGLU 2016 die Vergleichsgruppen der Teilnehmerstaaten der EU (VG EU) und die der OECD (VG OECD) zur Verfügung.

Die Darstellung der Ergebnisse in den nachfolgenden Kapiteln orientiert sich an den vorgestellten Entscheidungen: In Kapitel 3 werden die Ergebnisse für alle an IGLU 2016 teilnehmenden Staaten und Regionen auf der Gesamtskala Lesen dargestellt. Die Benchmark-Teilnehmer werden graphisch am unteren

Ende der aufgelisteten Teilnehmer durch einen kleinen Absatz getrennt aufgelistet. Darunter findet sich auch Dänemark mit Jahrgangsstufe 3 als Benchmark-Teilnehmer der Studienkomponente PIRLS *Literacy* (siehe Abschnitt 4.2.1). Die Leistungswerte der Benchmark-Teilnehmer fließen aus den bereits genannten Gründen nicht in die Berechnung des internationalen Mittelwerts ein. Bei allen anderen Ergebnissen, die in Tabellen und Abbildungen zu IGLU 2016 dokumentiert werden, werden diejenigen Teilnehmer und Benchmark-Teilnehmer berichtet, auf die mindestens eins der folgenden drei Kriterien zutrifft: (1) Mitglied der EU, (2) Mitglied der OECD und/oder (3) im Vergleich zu Deutschland signifikant bessere oder nicht signifikant unterschiedliche Leistung auf der Gesamtskala Lesen. Bei der Dokumentation von Trends im internationalen Vergleich werden diejenigen Bildungssysteme berücksichtigt, die (1) zusätzlich zu IGLU 2016 an mindestens zwei weiteren IGLU-Zyklen teilgenommen haben und (2) zugleich zum Erhebungszeitpunkt von IGLU 2016 Mitglied der EU und/oder der OECD sind und/oder auf der Gesamtskala Lesen in IGLU 2016 signifikant bessere oder nicht signifikant unterschiedliche Leistungen im Vergleich zu Deutschland erzielt haben. Eine Ausnahme ist die Flämische Gemeinschaft in Belgien, die in Trenddarstellungen berücksichtigt wird, obwohl sie neben IGLU 2016 nur an einem weiteren Zyklus (IGLU 2006) teilgenommen hat (siehe Abschnitt 4.2.2). In keiner Trenddarstellung werden Mittelwerte der Vergleichsgruppen VG EU und VG OECD berichtet, da sich über die vier Zyklen die Zusammensetzungen der Vergleichsgruppen durch andere Studienteilnehmer und neue Mitgliedschaften geändert haben.

Literatur

Anderson, R. & Pearson, P. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. Pearson (Hrsg.), *Handbook of reading research* (S. 255–291). White Plains, NY: Longman.

Autorengruppe Bildungsberichterstattung (2016). *Bildung und Migration*. Zugriff am 18.11.2017 unter https://www.bildungsbericht.de/de/bildungsberichte-seit-2006/bildungsbericht-2016/pdf-bildungsbericht-2016/h_web2016.pdf

Baumert, J. (2016). Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung. *Zeitschrift für Erziehungswissenschaft*, *19* (1), 215–253.

Baumert, J. & Weiß, M. (2002). Föderalismus und Gleichwertigkeit der Lebensverhältnisse. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000 – Die Länder der Bundesrepublik im Vergleich* (S. 39–53). Opladen: Leske + Budrich.

BMBF – Bundesministerium für Bildung und Forschung (Hrsg.). (2001). *TIMSS – Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente.* Bonn: BMBF Publik.

Boomsma, A., van Duijn, M. A. J. & Snijders, T. A. B. (Hrsg.). (2000). *Essay on item response theory*. New York: Springer.

Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C. & Walther, G. (Hrsg.). (2008). *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M., Schwippert, K. & Valtin, R. (Hrsg.). (2007). *IGLU 2006: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R. & Walther, G. (Hrsg.). (2004). *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich.* Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (Hrsg.). (2003). *Erste Ergebnisse aus IGLU. Schulleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Valtin, R. & Walther, G. (Hrsg). (2005). *IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien*. Münster: Waxmann.

Bos, W., Postlethwaite, T. N. & Gebauer, M. M. (2010). Potentiale, Grenzen und Perspektiven internationaler Schulleistungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 275–295). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (Hrsg.). (2012). *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Wendt, H., Köller, O. & Selter, C. (Hrsg.). (2012). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Campell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O. & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001* (2. Aufl.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Chall, J. (1983). *Stages of reading development*. New York: McGraw-Hill.

Christmann, U. & Groeben, N. (2001). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (S. 145–207). Baltmannsweiler: Schneider Hohengehren.

Ehmke, T., Klieme, E. & Stanat, P. (2013). Veränderungen der Lesekompetenz von PISA 2000 und PISA 2009. Die Rolle von Unterschieden in den Bildungswegen und in der Zusammensetzung der Schülerschaft. *Zeitschrift für Pädagogik*, *59*, 132–150.

Elley, W. B. (1992). *How in the world do students read*? *IEA study of reading literacy*. Den Haag: IEA.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Oxford: Pergamon.

Foy, P., Brossman, B. & Galia, J. (2011). Scaling TIMSS and PIRLS 2011 achievement data. In M. O. Martin & I. V. S. Mullis (Hrsg.), *TIMSS and PIRLS methods and procedures*. Zugriff am 17.10.2016 unter http://timssand pirls.bc.edu/methods/index.html

Heller, K. A. & Perleth, C. (2000). *KFT 4–12+R. Kognitiver Fähigkeitstest für 4. bis 12. Klasse, Revision*. Göttingen: Beltz Test.

Hooper, M., & Fishbein, B. (2017). Developing the PIRLS 2016 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Hrsg.), *Methods and procedures in PIRLS 2016.* Zugriff am 18.11.2017 unter https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-2.html

Hooper, M., Mullis, I. V. S. & Martin, M. O. (2015). PIRLS 2016 context questionnaire framework. In I. V. S. Mullis & M. O. Martin (Hrsg.), *PIRLS 2016 assessment framework* (2. Auflage, S. 31–54). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Howie, S. & Plomp, T. (2005). International comparative studies of education and large-scale change. In N. Bascia, A. Cumming, A. Datnow, K. Leithwood & D. Livingstone (Hrsg.), *International handbook of educational policy.* (Springer international handbooks of education, Bd. 13, S. 75–99). Berlin: Springer.

Hußmann, A., Wendt, H., Bos, W. & Rieser, S. (Hrsg.). (2018, i. Vorb.). *IGLU 2016. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente und Arbeit mit den Datensätzen.* Münster: Waxmann.

Joncas, M. & Foy, P. (2011). Sample design in TIMSS and PIRLS. In M. O. Martin & I. V. S. Mullis (Hrsg.), *TIMSS and PIRLS methods and procedures*. Zugriff am 16.11.2012 unter http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

IEA – International Association for the Evaluation of Educational Achievement. (2012). *Studies*. Zugriff am 16.11.2012 unter http://www.iea.nl/studies.html

Kasper, D., Wendt, H., Bos, W. & Köller, O. (2016). Trends in mathematischen und naturwissenschaftlichen Kompetenzen am Ende der Grundschulzeit in Deutschland. In H. Wendt, W. Bos, C. Selter, O. Köller, K. Schwippert & D. Kasper (Hrsg.), *TIMSS 2015. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 367–382). Münster: Waxmann.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessments. In J. Sabatini, E. Albro & T. O'Reilly (Hrsg.), *Measuring up: Advances in how to assess reading ability* (S. 21–37). Plymouth: Rowman & Littlefield Publishers.

Kintsch, W. (2013). Revisiting the construction-integration model of text comprehension and its implications for instruction. In D. Alvermann, N. Unrau & R. Ruddell (Hrsg.), *Theoretical models and processes of reading* (S. 807–841). Newark, DE: International Reading Association.

Klieme, E. (2013). Bildung unter undemokratischem Druck? Anmerkungen zur Kritik der PISA-Studie. In S. Lin-Klitzing, D. Di Fuccia & G. Müller-Frerich (Hrsg.), *Zur Vermessung von Schule* (S. 37–51). Bad Heilbrunn: Klinkhardt.

Klieme, E. & Vieluf, S. (2013). Schulische Bildung im internationalen Vergleich. Ein Rahmenmodell für Kontextanalysen in PISA. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung* (Zeitschrift für Pädagogik. Beiheft 59, S. 229–246). Weinheim: Beltz.

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* (Beschluss der 350. Kultusministerkonferenz vom 11.06.2015). Zugriff am 17.10.2016 unter https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsmonitoring.htm

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18* (1), 1–11.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking. Methods and practices* (2. Aufl.). New York: Springer.

Lankes, E.-M., Bos, W., Mohr, I., Plaßmeier, N., Schwippert, K., Sibberns, H. & Voss, A. (2003). Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung (IGLU) und ihrer Erweiterung um Mathematik und Naturwissenschaften (IGLU-E). In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 7–28). Münster: Waxmann.

Lehmann, R. H., Peek, R., Pieper, I. & Stritzky, R. v. (1995). *Leseverständnis und Lesegewohnheiten deutscher Schüler und Schülerinnen*. Weinheim: Beltz.

Lenhard, W., Lenhard, A. & Schneider, W. (2017). *ELFE II. Ein Leseverständnistest für Erst- bis Siebtklässler – Version II*. Göttingen: Hogrefe.

Linden, W. v. d. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Martin, M. O., Mullis, I. V. S. & Foy, P. (2015). Assessment design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I. V. S. Mullis & M. O. Martin (Hrsg.), *PIRLS 2016 assessment framework* (2. Auflage, S. 55–70). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P. & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J. & Kennedy, A. M. (2003). *Trends in children's reading literacy achievement 1991–2001: IEA's study of trends in reading literacy achievement in primary school in nine countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S. & Hooper, M. (Hrsg.). (2017). *Methods and procedures in PIRLS 2016*. Zugriff am 18.11.2017 unter https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. Hoboken, NJ: Wiley.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimation population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133–161.

Mullis, I. V. S., Kennedy, A. M., Martin, M. O. & Sainsbury, M. (Hrsg.). (2006). *PIRLS 2006. Assessment framework and specifications* (2. Aufl.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O. & Sainsbury, M. (2015). PIRLS 2016 reading framework. In I. V. S. Mullis & M. O. Martin (Hrsg.), *PIRLS 2016 assessment framework* (2. Auflage, S. 11–30). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S. & Martin, M. O. (Hrsg.). (2015). *PIRLS 2016 assessment framework.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P. & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Goh, S. & Prendergast, C. (Hrsg.). (2017). *PIRLS 2016 Encyclopedia: Education policy and curriculum in reading*. Zugriff am 18.11.2017 unter http://timssandpirls.bc.edu/pirls2016/encyclopedia/

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M. & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary school in 40 countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L. & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Hrsg.), *Methods and Procedures in PIRLS 2016* (S. 1.1–1.29). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Zugriff am 20.11.2017: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-1.html

Muraki, E. & Bock, D. (1999). *PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data* [Computer software]. Chicago, IL: Scientific Software,

OECD – Organisation for Economic Co-operation and Development. (2010). *PISA 2009 assessment framework. Key competencies in reading, mathematics and science*. Paris: OECD.

Prenzel, M. & Doll, J. (2002). Bildungsqualität von Schulen: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen [Beiheft]. *Zeitschrift für Pädagogik*, *45*.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. London: Sage Publication.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley + Sons.

Ruddell, R. & Unrau, N. (Hrsg.). (2004). Read as a meaning-construction process: The reader, the text, and the teacher. In R. Ruddell & N. Unrau (Hrsg.), *Theoretical models and processes of reading* (5. Auflage, S. 1462–1521). Newark, DE: International Reading Association.

Rumelhart, D. (1985). Toward an interactive model of reading. In H. Singer & R. Ruddell (Hrsg.), *Theoretical models and the processes of reading* (3. Auflage, S. 722–750). Newark, DE: International Reading Association.

Sheehan, K. M. (1985). *M-GROUP. Estimation of group effects in multivariate models* [Computer software]. Princeton, NJ: Educational Testing Service.

Statistisches Bundesamt (2017). *Bildung und Kultur. Allgemeinbildende Schulen*. Fachserie 11, Reihe 1, 2016/2017. Zugriff am 18.11.2017 unter https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100177004.pdf;jsessionid=29D3FD1BA09C7F36C1A0202842B06C01.InternetLive1?__blob=publicationFile

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (Hrsg.). (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich.* Münster: Waxmann.

UIS – UNESCO Institute for Statistics. (2015). *ISCED 2011 operational manual. Guidelines for classifying national education programmes and related qualifications*. Zugriff am 03.08.2016 unter http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

Tarelli, I., Wendt, H., Bos, W. & Zylowski, A. (2012). Ziele, Anlage und Durchführung der Internationalen Grundschul-Lese-Untersuchung (IGLU 2011). In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27–67). Münster: Waxmann.

Van Damme, J. & Bellens, K. (2016). Countries strive towards more quality and equity in education: Do they show success or failure? Evidence from TIMSS 2003 and 2011, for Grade 4. In M. Rosén, K. Yang Hansen & U. Wolff (Hrsg.), *Cognitive abilities and educational outcomes. A Festschrift in honour of Jan-Eric Gustafsson* (S. 127–148). Cham: Springer.

Wang, M., Haeterl, G. & Walberg, H. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63* (3), 249–294.

Weinert, F. E. & Helmke, A. (Hrsg.). (1997). *Entwicklung im Grundschulalter.* Weinheim: Beltz.

Wendt, H., Bos, W., Kasper, D., Walzebug, A., Goy, M. & Jusufi, D. (2016). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2015). In H. Wendt, W. Bos, C. Selter, O. Köller, K. Schwippert & D. Kasper (Hrsg.). *TIMSS 2015. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland* (S. 31–78). Münster: Waxmann.

Wendt, H., Bos, W., Selter, C., Köller, O., Schwippert, K. & Kasper, D. (Hrsg.). (2016). *TIMSS 2015. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Wendt, H., Bos, W., Tarelli, I., Vaskova, A. & Walzebug, A. (Hrsg.). (2016). *IGLU & TIMSS 2011. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente und Arbeit mit den Datensätzen*. Münster: Waxmann.

Wendt, H., Stubbe, T. C., Schwippert, K. & Bos, W. (Hrsg.). (2015). *10 Jahre international vergleichende Schulleistungsforschung in der Grundschule. Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011*. Münster: Waxmann

Wendt, H., Tarelli, I., Bos, W., Frey, K. & Vennemann, M. (2012). Ziele, Anlage und Durchführung der Trends in International Mathematics and Science Study (TIMSS 2011). In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 27–68). Münster: Waxmann.

Wendt, H., Walzebug, A., Bos, W., Smith, D. S. & Bremerich-Vos, A. (2017). Germany. In I. V. S. Mullis, M. O. Martin, S. Goh & C. Prendergast (Hrsg.), *PIRLS 2016 encyclopedia: Education policy and curriculum in reading*. Zugriff am 18.11.2017 unter http://timssandpirls.bc.edu/pirls2016/encyclopedia/countries/germany/

Werner, J. (1997). *Lineare Statistik*. Weinheim: Beltz.

Yu, A. & Ebbs, D. (2011). Translation and translation verification. In M. O. Martin & I. V. S. Mullis (Hrsg.), *TIMSS and PIRLS methods and procedures*. Zugriff am 16.11.2012 unter http://timssandpirls.bc.edu/methods/pdf/TP_ Translation_Verif.pdf