PISA-Studie 2012

Internationale Schulleistungsstudie der OECD

Seminararbeit im Rahmen des Bachelor-Seminars
"Ausgewählte Aspekte der Wirtschafts- und Sozialstatistik"
eingereicht von Jenny Steindl
LMU, Institut für Statistik

Seminarleiter: Prof. Dr. Thomas Augustin
Seminarbetreuerin: Eva Endres

München, den 22.09.2015

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Ein treibender Faktor der wirtschaftlichen Stärke der Bundesrepublik Deutschland ist dessen Innovationskraft, insbesondere da kaum auf natürliche Ressourcen zurückgegriffen werden kann. Hierbei belegt Deutschland mit mehr als 32.000 europäischen Patentanmeldungen im Jahr 2014 innerhalb der Europäischen Union den Spitzenplatz (Spiegel, 2014) und verdeutlicht die Relevanz gut ausgebildeter Fachkräfte. Umso schockierender war es für die breite Öffentlichkeit, als die erste PISA-Studie im Jahr 2000 große Defizite im deutschen Bildungswesen aufdeckte (vgl. FAZ a, 2013). Hierbei zeigte sich, dass 15-jährige deutsche Schüler im Vergleich zu ihren Altersgenossen in anderen Staaten oftmals große Rückstände aufweisen.
Als Konsequenz der Ergebnisse der PISA-Studie im Jahr 2000 wurden teilweise tiefgreifende Reformen im Bildungssystem eingeleitet. Exemplarisch kann hier die Vereinheitlichung der Bildungsstandards der Bundesländer genannt werden (vgl. z.B. FAZ b, 2013). Durch die wiederholte Erhebung der PISA Studie, welche im Jahr 2012 das fünfte Mal seit der Initiierung im Jahr 2000 durchgeführt wird, ist es möglich, den Erfolg der ergriffenen Bildungsreformen zu kontrollieren (KMK, 2010). Hierbei zeigt sich in der Tat eine stetige Verbesserung der deutschen Schüler im Zeitverlauf (eine kompakte Zusammenfassung bietet FAZ b, 2013).
Die PISA-Studie selbst ist nicht nur durch ihren Umfang von mehr als 500.000 Schülern (Sälzer & Prenzel, 2013, S. 21) eine sehr aufwändige Studie, sondern sie geht auch mit teils sehr komplexen statistischen Methoden und Auswertungen einher. Gerade diese Komplexität hat oftmals eine richtige Einordnung der Ergebnisse innerhalb des öffentlichen Diskurses erschwert.
Folglich legt diese Arbeit einen Schwerpunkt auf die Methodik der PISA-Studie des Jahres 2012. Hierbei gibt Kapitel 2 grundlegende Informationen zur PISA-Studie 2012 während Kapitel 3 die Methodik erläutert, insbesondere die Stichprobenziehung, die Gewichtung der Stichprobe sowie die zur Auswertung verwendeten Modelle. Schließlich erläutert Kapitel 4 die Skalierung und Interpretation der Modellergebnisse während in Kapitel 5 die wichtigsten Ergebnisse der PISA-Studie des Jahres 2012 zusammengefasst werden.

2 Überblick über PISA 2012

Bevor die Methodik der PISA-Studie genauer beleuchtet wird ist es sinnvoll sich mit den grundlegenden Sachverhalten bezüglich PISA zu beschäftigen, wie etwa den Teilnehmern oder den Zielen der Studie. Diesbezüglich geben die beiden folgenden Kapitel einen Überblick.

2.1 Allgemeine Beschreibung

Das *"Programme for International Student Assessment"* ist eine Studie der *"Organisation for Economic Co-Operation and Development"* (OECD). Im Jahr 2012 haben insgesamt 65 Staaten an der PISA-Studie teilgenommen. Dies entspricht gegenüber der ersten Durchführung im Jahr 2000 einer Steigerung um 33 Teilnehmer und zeugt von einem wachsenden regen internationalen Interesse an der Studie. Die 65 Teilnehmerstaaten im Jahr 2012 bestehen aus den 34 OECD-Mitgliedsstaaten sowie 31 Partnerstaaten (Sälzer & Prenzel, 2013). Diese werden in Grafik 1 zusammengefasst:

Albanien*	Jordanien*	Republik Serbien*
Argentinien*	Kanada	Rumänien*
Australien	Kasachstan*	Russische Föderation*
Belgien	Katar*	Schweden
Brasilien*	Kolumbien*	Schweiz
Bulgarien*	Korea	Shanghai (China)*
Chile	Kroatien*	Singapur*
Chinesisch Taipeh*	Lettland*	Slowakische Republik
Costa Rica*	Liechtenstein*	Slowenien
Dänemark	Litauen*	Spanien
Deutschland	Luxemburg	Thailand*
Estland	Macau (China)*	Tschechische Republik
Finnland	Malaysia*	Tunesien*
Frankreich	Mexiko	Türkei
Griechenland	Neuseeland	Ungarn
Hongkong (China)*	Niederlande	Uruguay*
Indonesien*	Norwegen	Vereinigte Arabische Emirate*
Irland	Österreich	Vereinigte Staaten
Island	Peru*	Vereinigtes Königreich
Israel	Polen	Vietnam*
Italien	Portugal	Zypern*
Japan	Republik Montenegro*	

Anmerkung: * OECD-Partnerstaaten

Fig. 1: An PISA 2012 teilnehmende Staaten (Sälzer & Prenzel, 2013, S. 22)

Die Studie wird in einem Zyklus von drei Jahren wiederholt, wobei jedes Mal die Kompetenzbereiche Lesen, Mathematik und Naturwissenschaften im Fokus stehen. Dabei wird bei jeder Durchführung wechselnd ein einzelner Kompetenzbereich besonders intensiv betrachtet. Bei der ersten PISA Erhebung im Jahr 2000 lag der Schwerpunkt der Studie auf dem Kompetenzbereich Lesen. 2012 wurde Mathematik als Schwerpunkt der PISA-Studie gesetzt (Sälzer & Prenzel, 2013).

Primäre Aufgabe der PISA-Studie ist es, den OECD-Mitgliedsstaaten Daten vorzulegen, die zu politisch-administrativen Entscheidungen zur Verbesserung der nationalen Bildungssysteme beitragen können. Das zyklische Programm der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung wird von allen Mitgliedsstaaten gemeinschaftlich getragen und verantwortet. Dabei sollen die Leistungen 15-jähriger Schülerinnen und Schüler in den verschiedenen Kompetenzbereichen erfasst werden. National und international wird somit die Leistung der Zielpopulation aufgezeigt, um eventuelle Defizite aufzudecken und gegebenenfalls über die nationalen Bildungssysteme korrektive Maßnahmen zu ergreifen. Gemäß einer Vereinbarung zwischen dem Bundesministerium für Bildung und Forschung und der Ständigen Konferenz der Kultusminister der Länder ist die Bundesrepublik Deutschland an dieser Stelle an der PISA-Studie beteiligt (Sälzer & Prenzel, 2013).
Insgesamt wurden bei der PISA-Studie 2012 in den 65 teilnehmenden Staaten etwa 500.000 Schüler auf ihre Leistungsfähigkeit getestet. Die ausgewählten Schüler stehen stellvertretend für die Grundgesamtheit von insgesamt rund 28 Millionen Fünfzehnjährigen in den teilnehmenden Ländern. In der Bundesrepublik Deutschland wurden hierzu aus insgesamt 247 Schulen 5000 Schüler für die Teilnahme an der Studie ausgewählt (Sälzer & Prenzel, 2013).

2.2 Durchführung

Die Erhebung der PISA-Studie 2012 fand im Zeitraum 01. März bis 30. Juni 2012 statt. Die Schülerinnen und Schüler hatten zur Bearbeitung der Tests jeweils zwei Stunden Zeit, wobei die Tests an zwei Tagen durchgeführt wurden. Eine Ausnahmeregelung wurde hierbei an Sonderschulen getroffen, an welchen die Schüler eine verkürzte Testversion bekamen und diese an einem Tag bearbeiten sollten (Sälzer & Prenzel, 2013).

Den Schülern wurden vorab Instruktionen zur Durchführung der Tests aus einem Skript von geschulten Testleiterinnen und Testleitern vorgelesen, um die Vergleichbarkeit der Durchführungsbedingungen zu gewährleisten. Eine nachträgliche Testerhebung wurde an einer Schule gestattet, da am Tag der Testdurchführung vier oder mehr Schüler fehlten, die zuvor für die Studie ausgewählt wurden. Um den Datenschutz zu gewährleisten wurden die Datenschutzbeauftragten der Länder in die Vorbereitung und Durchführung der Studie mit einbezogen. Zudem wurden zur Konzipierung der Tests international führende Institutionen und Experten zur Testentwicklung hinzugerufen um sicher zu stellen, dass dieser geeignet ist, die gesetzten Anforderungen und Ziele der Studie zu erreichen (Sälzer & Prenzel, 2013).

Eltern und Schüler wurden vorab über das Vorgehen sowie die Ziele der Studie in Kenntnis gesetzt und erklärten sich mit der Teilnahme schriftlich einverstanden. Um eine ordnungsgemäße Durchführung der Tests zu gewährleisten fanden in Deutschland an insgesamt 35 Schulen unangemeldete Kontrollen statt, um zu bestätigen, dass die Untersuchungsbedingungen eingehalten wurden (Sälzer & Prenzel, 2013).

3 Methodik

Ein elementarer Baustein der PISA-Studie ist die hierfür verwendete Methodik. Diese beinhaltet sowohl die Stichprobenziehung, deren Gewichtung, als auch die für die Schätzung der zu untersuchenden Populationsparameter verwendeten Modelle. Diese Punkte werden in den folgenden Kapiteln detaillierter betrachtet.

3.1 Stichprobenerhebung

Die international vorgegebene Zielpopulation, d.h. die Grundgesamtheit, besteht aus allen fünfzehnjährigen Schülern, wobei die genaue Altersdefinition mit dem internationalen PISA-Konsortium abgestimmt wurde. Da die Grundgesamtheit etwa 28 Millionen 15-Jährige umfasst, ist eine Vollerhebung aus organisatorischen, zeitlichen und finanziellen Gründen nicht möglich. Aus diesem Grund wurde eine systematische Teilerhebung durchgeführt, welche durch die Ziehung einer Zufallsstichprobe erfolgte. Dabei soll gewährleistet werden, dass Rückschlüsse von der Stichprobe auf die Grundgesamtheit zugelassen werden können. Da Schulleistungsstudien, insbesondere Large-Scale-Assessments besondere Anforderungen stellen, werden im Allgemeinen zwei- oder mehrstufige Auswahlverfahren angewendet (Heine et al., 2013).

Stichprobenplan und Ziehung der Stichprobe

Im Detail setzt sich die Grundgesamtheit der PISA Studie aus der Kohorte der 15-jährigen Schüler ab der siebten Klasse zusammen. Das sind in Deutschland genau die Schüler, die zwischen dem 01. Januar und dem 31. Dezember 1996 geboren wurden und mindestens in der siebten Klasse sind. Die Auswahl der teilnehmenden Schüler erfolgt anhand eines zweistufigen Verfahrens. In Deutschland wurde ein erweitertes zweistufiges Verfahren angewandt, welches in den nachfolgenden Abschnitten detaillierter betrachtet wird (Heine et al., 2013).

Das grundlegende Verfahren besteht darin, dass zunächst eine Zufallsstichprobe von Schulen gezogen wird. Aus diesen wird zufällig eine Gruppe von 25 Schülern in der relevanten Altersklasse ausgewählt. Da in Deutschland ein erweitertes Verfahren zur Anwendung kam, wurden zudem zwei vollständige neunte Klassen per Zufallsverfahren ausgewählt, die an der PISA-Studie teilnahmen. Zusätzlich beteiligte sich Deutschland an dem Computer-based Assessment (CBA). Für den CBA standen die Kompetenzbereiche Lesen, Mathematik und Problemlösung zur Auswahl, wobei sich

die Teilnehmerstaaten wahlweise auf einen Kompetenzbereich konzentrieren mussten. Deutschland fokussierte sich auf den Bereich Problemlösung. Für das CBA wurden aus den pro Schule ausgewählten 25 Schülern noch einmal zufällig 14 Schüler separiert (Heine et al., 2013).

Wie diese mehrstufige Auswahl für Deutschland im Detail erfolgte wird in den nächsten Abschnitten genauer erläutert.

1. Schritt: Erstellung eines Sampling-Frame

Im ersten Schritt wurde die für Deutschland zu analysierende Grundgesamtheit bestimmt. Da bei der späteren Stichprobenziehung zunächst die Schulen ausgewählt wurden war es notwendig, die Grundgesamtheit aller Schulen zu bestimmen, welche potentiell von 15-jährigen Schülern besucht wurden. Diese Grundgesamtheit wird als Sampling-Frame bezeichnet. Hierbei ist es wichtig zu erwähnen, dass bei dem Sampling-Frame keine Schulen mehrfach, fehlerhaft oder unvollständig dokumentiert werden durften (Heine et al., 2013). Eine Einrichtung wird gemäß der Richtlinie des KMK (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) als Schule gewertet, wenn diese hinsichtlich des Lehrplans und des Qualifikationsniveaus als eigenständige Einheit verstanden werden kann (vgl. KMK, 2013, S.8). Dies bedeutet, dass sich die Definition einer Schule nicht aus der räumlichen Gegebenheit ergibt. Um das Sampling-Frame aller relevanten Schulen zu erstellen, wurden von den 14 statistischen Landesämtern folgende Informationen auf Basis der Schulstatistik 2010/2011 eingeholt:

- die offizielle Schulnummer
- die Schulart
- die Anzahl an Schüler in den Geburtsjahrgängen 1993 bis 1996
- die Anzahl der Schüler in den 8. bis 10. Klassen
- die Anzahl der Klassen 8. bis 10.
- Informationen über Schulveränderungen (Schließungen, Zusammenlegungen und Schulart)
- für Förderschulen die Informationen über die Förderschwerpunkte (Heine et al., 2013, S. 314)

2. Schritt: Ziehung der Schulstichprobe

Nachdem die Grundgesamtheit aller relevanten Schulen erstellt wurde, konnte diese anhand mehrerer sogenannter Stratifizierungskriterien genauer unterteilt werden. Diese Unterteilung bietet mehrere Vorteile (einen detaillierten Überblick über die Stratifizierung bieten Kalton, 1983; Kish, 1995; Levy & Lemeshaw, 2008 und Daniel, 2012):

Erfolgt die Unterteilung des Sampling Frames derart, dass die einzelnen Untergruppen in sich möglichst homogen sind, so kann die Effizienz der Stichprobe und der daraus abgeleiteten Analysen erhöht werden. Dies bedeutet, dass die Populationsparameter verlässlicher geschätzt werden. Das ermöglicht es, mit kleineren Stichproben verlässliche Ergebnisse zu erhalten. Dies lässt sich am besten anhand der folgenden Überlegung verdeutlichen: Ziel der Stratifizierung ist es, Untergruppen zu bilden, die in sich möglichst homogen sind. Gleichzeitig sollen sich die Untergruppen voneinander möglichst unterscheiden. Dementsprechend sind sich die Individuen innerhalb eines Stratums hinsichtlich der zu untersuchenden Merkmale recht ähnlich. Daraus folgt unmittelbar, dass die Stichprobenvarianz der untersuchten Merkmale innerhalb eines Stratums geringer ist, als dies in der gesamten Stichprobe der Fall wäre. Durch die Stratifizierung wird somit die Varianz der gesamten Stichprobe auf die Variationen zwischen den einzelnen Gruppen konzentriert, sodass diese in den einzelnen Gruppen deutlich geringer ausfällt. Zudem ermöglicht die Stratifizierung zu überprüfen, ob die Charakteristiken der Stichprobe mit jenen der Grundgesamtheit übereinstimmen. Darüber hinaus kann die Analyse unterschiedlich nach bestimmten Kriterien erfolgen. So können beispielsweise mit Hilfe der Stratifizierung die Ergebnisse einer Hauptschule mit den Ergebnissen einer Realschule verglichen werden (Heine et al., 2013).

Stratifizierungen können nach expliziter Stratifizierung und impliziter Stratifizierung unterschieden werden. Bei der expliziten Stratifizierung wird die Grundgesamtheit zunächst anhand bestimmter Charakteristiken in unabhängige Gruppen unterteilt. Aus jeder Gruppe wird anschließend eine separate Stichprobe gezogen. Bei der impliziten Stratifizierung hingegen werden die anhand der expliziten Strata gezogenen Stichproben in einzelne Merkmalsklassen unterteilt (Heine et al., 2013).

In Deutschland wird die Grundgesamtheit aller Schulen mittels zweier expliziter und zweier impliziter Stratifizierungsverfahren aufgeteilt:

Die Schulen werden zunächst in allgemeinbildende Schulen, Förderschulen und Berufsschulen unterteilt, was einem expliziten Stratifizierungsverfahren entspricht. Anschließend folgt eine weitere explizite Stratifizierung, indem die allgemeinbildenden Schulen nach Bundesländern aufgeteilt werden. Für die implizite Stratifizierung werden die Berufs- und Förderschulen nach Bundesländern aufgeteilt. Zudem werden die allgemeinbildenden Schulen in Hauptschulen, Realschulen und Gymnasien unterteilt. Daraus folgt, dass die Schulen in 18 explizite Strata aufgeteilt werden (16 Bundesländer und 2 Förder- und Berufsschulen) aus denen jeweils separat zufällige Stichproben gezogen werden (Heine et al., 2013).
Die Stichproben aus den jeweiligen Strata wurden vom internationalen PISA Konsortium gezogen. Es ergab sich ein Umfang von insgesamt 247 in die Stichprobe gezogenen Schulen. Hierbei wurde darauf geachtet, dass der proportionale Anteil der Schüler in den Stichproben approximativ jenem der Grundgesamtheit innerhalb der Staaten gleicht (Heine et al., 2013). Die nachfolgende Abbildung gibt einen Überblick über die Anzahl der in Deutschland ausgewählten Schulen, aufgeteilt nach Schulart und Bundesländern:

	Hauptschule	Integrierte Gesamtschule	Schule mit mehreren Bildungsgängen	Realschule	Gymnasium	Förderschule	Berufsschule	Gesamt
Baden-Württemberg	2	1	5	13	12	2	4	**39**
Bayern	10	1	0	13	12	1	6	**43**
Berlin	0	0	4	0	3	0	0	**7**
Brandenburg	0	0	2	0	2	1	0	**5**
Bremen	0	0	1	0	1	0	0	**2**
Hamburg	0	0	2	0	2	0	1	**5**
Hessen	2	3	0	5	6	1	1	**18**
Mecklenburg-Vorpommern	0	0	2	0	1	0	0	**3**
Niedersachsen	4	2	0	9	8	2	1	**26**
Nordrhein-Westfalen	10	10	0	15	18	3	2	**58**
Rheinland-Pfalz	0	2	4	1	4	1	0	**12**
Saarland	0	0	1	0	1	0	1	**3**
Sachsen	0	0	3	0	3	1	1	**8**
Sachsen-Anhalt	0	0	2	0	2	0	0	**4**
Schleswig-Holstein	0	4	1	0	3	1	0	**9**
Thüringen	0	0	3	0	1	0	1	**5**
Gesamt	28	23	30	56	79	**13**	**18**	247

Fig. 2: In die Stichprobe gezogene Schulen nach Bundesland und Schulart (Heine et al., 2013, S. 316)

3. Schritt: Ziehung der Schülerstichprobe

Bevor die Schülerstichprobe gezogen werden konnte, mussten bestimmte Informationen der einzelnen Schüler in einer Liste eingetragen werden. Diese Informationen bestanden aus dem Vor- und Nachnamen, dem Geschlecht, dem Geburtsjahr und Geburtsmonat, der Klassenbezeichnung und der Information über einen möglichen Förderbedarf der Schüler. Diese Liste wurde mit der Liste des Statistischen Landesamtes verglichen. Um den Datenschutz einhalten zu können, wurden den Schülern Pseudonyme zugeschrieben. Für weitere Details hierzu sei auf die Seminararbeit "Anonymisierungsverfahren" hingewiesen. Waren keine Diskrepanzen zwischen den Schülerlisten der Schulen und der des Statistischen Landesamtes erkennbar, wurde mit der Ziehung der Stichprobe begonnen (Heine et al., 2013).

Aus den vorher ausgewählten allgemeinbildenden Schulen wurde nun mittels eines Zufallsverfahrens eine Stichprobe von 25 Schülern pro Schule gezogen. Zudem wurde an den allgemeinbildenden Schulen eine Stichprobe von zwei 9. Klassen gezogen, die vollständig mit in die Stichprobe aufgenommen wurden. Zusätzlich wurden aus den 25 gezogenen Schülern nochmals 14 Schüler gezogen, die an dem Testprogramm CBA teilnahmen. Bei allen ausgewählten Berufs- und Förderschulen konnte eine Vollerhebung aller neunten Klassen durchgeführt werden. Dieses Verfahren soll durch die unten stehende Grafik veranschaulicht werden (Heine et al., 2013).

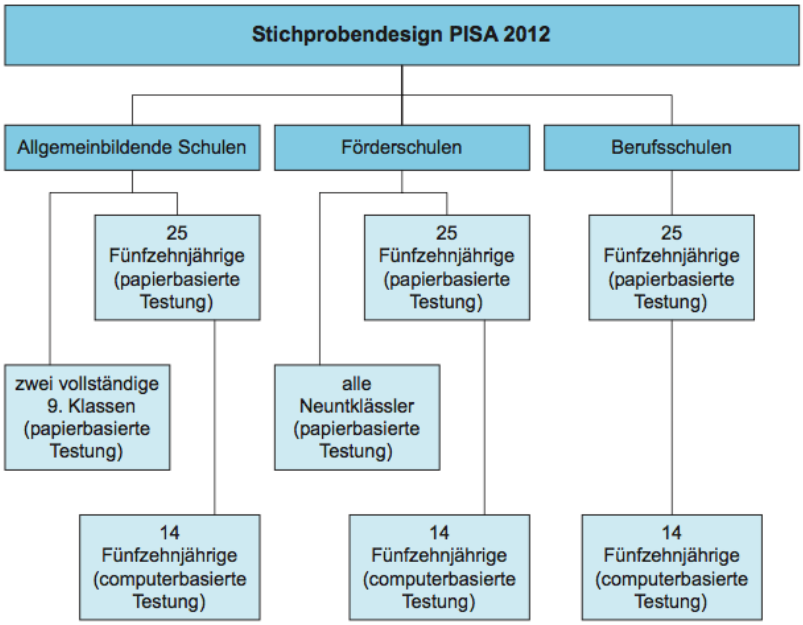

Fig. 3: Stichprobendesign PISA 2012 (Heine et al., 2013, S. 317)

Realisierte Stichproben

In den vorherigen Abschnitten wurden die Grundgesamtheit, die Schulstichproben und Schülerstichproben genauer erläutert. Nun gilt es, eine weitere Unterteilung zu betrachten. In diesem Abschnitt wird zwischen Brutto- und Nettostichproben unterschieden. Unter Bruttostichproben versteht man alle für die PISA-Studie 2012 ausgewählten Schülerinnen und Schüler, wohingegen man bei der Nettostichprobe nur von der Anzahl von Schülern ausgeht die tatsächlich an der Studie teilnahmen (Heine et al., 2013).

Von den 247 in der Stichprobe befindlichen Schulen sprachen sich fünf Privatschulen gegen eine Teilnahme an der PISA-Studie aus. Da es den Schulen freigestellt ist an der Studie teilzunehmen, fielen diese fünf Schulen aus der Stichprobe heraus. War dies der Fall, so konnten Ersatzschulen für die entfallenen Schulen bestimmt werden (Heine et al., 2013). Des Weiteren gab es 17 Ausfälle von Schulen, für die keine Ersatzschulen bestimmt werden konnten, da diese entweder nicht vorgesehen waren oder es keine gab. Unter diesen 17 Schulen befanden sich 11 Berufsschulen, die keine 15- jährigen Schüler unterrichteten. Weitere 2 Privatschulen entfielen, die selbst schon als Ersatzschule vorgesehen waren. Zwei der gezogenen Schulen wurden aufgelöst und existierten somit nicht mehr. Für eine ausgefallene Privatschule wurde kein Ersatz gefunden. Eine weitere dänische Schule fiel aus der gezogenen Schulstichprobe, da sie nur vier 15-jährige Schüler an der Schule hatten (Heine, et al., 2013). Somit verblieben 230 Schulen, die an der PISA-Studie 2012 teilnahmen. Bei den Schulen wurde eine Teilnehmerquote von $98,3\%$ erreicht. Bei den einzelnen Schülern betrug die Quote der Teilnehmer $98,2\%$. Somit wurde die Mindestvorgabe der OECD Standards erfüllt. Die deutsche PISA-Studie umfasste $n = 5001$ Schüler und $n = 9998$ Neuntklässler (Heine, et al., 2013).

3.2 Gewichtung

Ein wichtiger Faktor um eine verlässliche Stichprobe zu bekommen ist die Gewichtung der Schüler in der Stichprobe, die an dem Test teilgenommen haben. Diese wird in den folgenden Abschnitten erläutert. Diese beziehen sich dabei auf den *Technical Report (PISA 2012)* der OECD (2014), Kapitel 8, S. 132ff.

Notwendigkeit der Gewichtung

Ein Umstand der bei der Analyse berücksichtigt werden muss ist, dass nicht alle Schüler dieselbe Wahrscheinlichkeit besitzen, zufällig in die Stichprobe aufgenommen zu werden. Zudem können bestimmte Subpopulationen über- oder unterrepräsentiert sein. Dies lässt sich auf verschiedene Ursachen zurückführen: Das Stichprobendesign kann beabsichtigt bestimmte Subpopulationen über- oder unterrepräsentieren. Eine Übergewichtung wird vollzogen, um etwa wichtige dennoch kleine Populationsgruppen effektiv analysieren zu können. Eine Untergewichtung bestimmter Gruppen hat meist das Motiv Kosten zu sparen. Ein weiterer Grund können falsche Angaben über die Größe der Schule sein. Sollte eine Schule fälschlicherweise eine hohe Schülerzahl angeben, wird planmäßig aus dieser Schule eine hohe Schülerzahl gezogen. Da die Schülerzahl in Wirklichkeit allerdings geringer ist, haben die Schüler der Schule eine höhere Wahrscheinlichkeit in die Stichprobe zu gelangen, als ihnen ursprünglich zugedacht war. Weiterhin kann eine teilweise oder gänzliche Nicht-Teilnahme einer Schule dazu führen, dass bestimmte Populationsgruppen in der Stichprobe über- oder unterrepräsentiert sind. Dasselbe kann dadurch verursacht werden, dass einzelne Schüler in der Stichprobe nicht am Test teilnehmen. Die Über- bzw. Unterrepräsentation von Subpopulationen kann zu Schätzern mit hohen Stichprobenfehlern und einer unkorrekten Abbildung in den nationalen Schätzungen führen. Um diesen Umstand entgegen zu wirken, werden die Schüler in der Stichprobe im Nachhinein gewichtet, um die negativen Effekte auszugleichen (OECD, 2014). Kish (1992) zeigt, dass die Gewichtung zwar geringfügig die Schätzer verzerren kann, jedoch die Standardfehler erheblich reduziert. Die für die Gewichtung verwendete Methodik geht auf Cochran (1977), Särndal et al. (1992) sowie Lohr (2010) zurück und wird in den nächsten Abschnitten ausführlicher erläutert.

Allgemeine Gewichtung

Die gesamte Gewichtung W_{ij}, die einem Schüler j aus der Schule i in der Stichprobe zugeteilt wird, ergibt sich aus zwei schulspezifischen Basisgewichten w_{1i} und w_{2ij} sowie fünf Adjustierungsgewichten t_{1i}, t_{2ij}, f_{1i}, f_{2ij} sowie f^A_{1ij}, so dass gilt:

$$W_{ij} = t_{2i} f_{1i} f_{2ij} f^A_{1ij} w_{2ij} w_{1i}$$

Die Bedeutung sowie Bestimmung der einzelnen Gewichte wird in den folgenden Abschnitten erläutert:

Das Schulbasisgewicht

Das Schulbasisgewicht w_{1i} ist generell das Reziproke der Wahrscheinlichkeit, dass eine Schule i in die Stichprobe aufgenommen wird. Das Schulbasisgewicht ist definiert zu:

$$w_{1i} = \begin{cases} I_g / MOS_i & , \quad falls < MOS_i < I_g \\ 1 & , \quad \text{sonst} \end{cases}$$

Dabei ist MOS_i die geschätzte Zahl an 15-jährigen Schülern an der Schule i in der Stichprobe. I_g ergibt sich aus dem Verhältnis zwischen der gesamten Zahl von 15-jährigen Schülern in einem bestimmten Stratum g und der Anzahl an Schulen die aus diesem Stratum ausgewählt werden (OECD, 2014, S. 133). Die Bedeutung des Schulbasisgewichtes lässt sich am besten durch ein Zahlenbeispiel verdeutlichen: Angenommen eine Schule i in der Stichprobe hat 100 15-jährige Schüler, so ist $MOS_i = 100$. Sei die Anzahl an 15-jährigen Schülern in einem Stratum $g = 1$ gleich 150.000 und es werden aus diesem Stratum 150 Schulen ausgewählt, so ist $I_1 = 150.000/150 = 1000$. Daraus folgt $w_{1i} = 1000/100 = 10$. Das bedeutet, dass die Schule i etwa 10 Schulen in der Population repräsentiert (OECD, 2014, S. 133).

Der Schulbasisgewicht Zuschnittfaktor t_{1i}

Der Schulbasisgewicht Zuschnittfaktor t_{1i} dient dazu, unerwartet hohe Werte des Schulbasisgewichts w_{1i} auszugleichen. Dieser ergibt sich aus dem Verhältnis zwischen dem zugeschnittenen und dem nicht zugeschnittenen Schulbasisgewicht. Der

Zuschnittfaktor t_{1i} kommt jenen Schulen zu, deren tatsächliche Schülerzahl deutlich höher liegt als jene, die zum Zeitpunkt der Stichprobenziehung angenommen wurde. Dies waren speziell jene Schulen, bei welchen die Zahl der 15-jährigen Schülern 3 x $max(TCS, MOS_i)$ überstieg. TCS (target cluster size) ist hierbei entweder 35 (für Schulen die nicht an der financial literary study teilnahmen) bzw. 43 (bei Schulen die an der financial literary study teilnahmen). Für diese Schulen wurde bei der Bestimmung des Schulbasisgewichts MOS_i nicht auf die ursprünglich angenommen Zahl an 15-jährigen Schülern festgesetzt, sondern auf 3 x $max(TCS,\ MOS_i)$ (OECD, 2014, S. 134).

Das schülerspezifische Schulbasisgewicht w_{2ij}

Das schülerspezifische Schulbasisgewicht w_{2ij} ist definiert als das Reziproke der Wahrscheinlichkeit, dass ein Schüler j aus der Schule i in die Stichprobe gelangt. Da innerhalb einer Schule jeder Schüler dieselbe Wahrscheinlichkeit besitzen soll, in die Stichprobe aufgenommen zu werden ergibt sich w_{2ij} zu:

$$w_{2ij} = \frac{enr_i}{sam_i}$$

enr_i ist die tatsächliche Zahl an 15-jährigen Schülern am Tag des Tests an der Schule i und kann somit vom MOS_i abweichen. sam_i bezeichnet die Zahl der Schüler, die an der Schule i für den Test ausgewählt werden sollen. Somit gilt $w_{2ij} \geq 1$. Das bedeutet, dass ein Schüler in der Stichprobe mindestens einen, wenn nicht mehrere Schüler an der Schule i repräsentiert (OECD, 2014, S. 134).

Der Adjustierungsfaktor für nicht teilnehmende Schulen f_{1i}

f_{1i} dient als Korrekturfaktor für eine Schule i, der die Nicht-Teilnahme einer Schule kompensieren soll, die in ihren Eigenschaften ähnlich zu Schule i ist. Der Korrekturfaktor kommt nur in solchen Fällen zur Anwendung, wenn für die nicht teilnehmende Schule kein Ersatz mehr bestimmt werden kann. Um diesem Umstand zu begegnen werden die Schulen zunächst anhand der Stratifizierungkriterien in Gruppen eingeteilt. Hierbei werden für jedes Land 10-15 Gruppen erstellt. Der Adjustierungsfaktor der einer Schule i aufgrund der Nicht-Teilnahme einer ähnlichen Schule zukommt ist definiert zu:

$$f_{1i} = \frac{\sum\limits_{k \in \Omega(i)} w_{1k} enr(k)}{\sum\limits_{k \in \Gamma(i)} w_{1k} enr(k)}$$

Der Zähler gibt hierbei die mit dem jeweiligen Schulbasisgewicht versehene Summe der tatsächlichen Schülerzahlen jener Schulen innerhalb einer Gruppe an, unabhängig davon ob die Schulen letztendlich am Test teilgenommen haben. Dies bedeutet, dass der Zähler sich auf alle Schulen innerhalb einer Gruppe bezieht, egal ob diese teilnahmen, nicht teilnahmen oder als Ersatz für nicht teilnehmende Schulen fungierten. Der Nenner bezieht sich hingegen auf die tatsächliche Zahl an Schülern innerhalb einer Gruppe, die an dem Test teilgenommen haben. Dies sind folglich die Schüler der Schulen, die ursprünglich für den Test vorgesehen waren und auch teilnahmen sowie jener Schulen, die als Ersatz für andere Schulen teilgenommen haben (OECD, 2014, S. 134f).

Die Adjustierung für abweichende Klassenstufen f_{1i}^{A}

Gelegentlich kam es vor, dass Schulen zwar an der PISA-Studie teilnahmen, jedoch nur jene 15-jährigen Schüler dafür auswählten, die auch in der für dieses Alter vorgesehenen Klassenstufen waren. Da durch dieses Auswahlverfahren die Stichprobe in ihren Charakteristiken nicht mit der Grundgesamtheit übereinstimmt, bedarf dieser Umstand einer Bereinigung. Dies geschieht durch den Adjustierungsparameter f_{1i}^{A}. Dieser ist wie folgt definiert:

$$f_{1i}^{A} = \begin{cases} \dfrac{\sum\limits_{k \in C(i)} w_{1k} enra(k)}{\sum\limits_{k \in B(i)} w_{1k} enra(k)} & \\ & , \text{ falls nicht in der vorhergesehenen Klasse} \\ 1 & , \text{ sonst} \end{cases}$$

Der Nenner bezieht sich auf all jene Schulen, bei denen alle 15-jährigen Schüler unabhängig von deren Klassenstufe für die PISA-Studie zur Verfügung standen. Er beschreibt die tatsächliche Zahl von 15-jährigen Schülern an diesen Schulen, die nicht in die für dieses Alter vorgesehene Klassenstufe besuchten. Der Zähler erweitert diese Gruppe um jene 15-jährigen Schüler, die Schulen angehören, welche nur die für dieses Alter vorgesehene Klassenstufen für den Test zuließen, und die nicht in eine ihres Alters entsprechende Klassenstufe besuchten. Die Schülerzahlen werden dabei jeweils

mit dem Schulbasisgewicht versehen (OECD, 2014, S. 135).

Der schülerspezifische Adjustierungsparameter für nicht teilnehmende Schüler f_{2i}

f_{2i} dient dazu, die Verzerrung durch nicht teilnehmende Schüler auszugleichen. Hierfür werden zunächst schulspezifische Cluster gebildet, welche die Schüler einer jeden Schule nach den Merkmalen *Geschlecht* und *Leistungsklasse* (niedrig/hoch) in vier Gruppen einteilen. Auf Basis dieser Gruppen ergibt sich f_{2i} somit zu:

$$f_{2i} = \frac{\sum\limits_{k \in X(i)} f_{1i} w_{1i} w_{2ik}}{\sum\limits_{k \in \Delta(i)} f_{1i} w_{1i} w_{2ik}}$$

Der Nenner bezieht sich auf alle Schüler innerhalb einer Gruppe, die an dem Test teilgenommen haben. Der Zähler umfasst die Gewichte jener Schüler, die tatsächlich am Test teilgenommen haben, als auch jener, die teilnehmen hätten sollen, dies jedoch nicht getan haben (OECD, 2014, S. 137).

Der Zuschnittfaktor für die Schüler t_{2ij}

Innerhalb eines jeden expliziten Stratums soll die Wahrscheinlichkeit in die Stichprobe zu gelangen für alle Schüler gleich sein. Da jedoch durch die bereits genannten Gründe wie der Nicht-Teilnahmen von Schülern oder Schulen dies oft nicht der Fall ist, muss dieser Umstand durch die Gewichtung behoben werden. In ungünstigen Fällen kann es jedoch vorkommen, dass einem Schüler ein sehr hohes Gewicht zukommt. Dies ist unvorteilhaft, da hohe Gewichte zu einem starken Anstieg der Stichprobenvarianz führen können. Das ist unter anderem dann der Fall, wenn jene Schüler ein sehr hohes Gewicht bekommen, deren Leistungen besonders stark vom (Teil-) Stichprobenmittel abweichen. Daher werden sehr hohe Gewichte reduziert. Speziell werden alle schülerspezifischen Gewichte die mehr als viermal größer als das Mediangewicht sind, auf das Vierfache des Mediangewichts beschränkt. Daher ist t_{2ij} definiert als das Verhältnis zwischen dem finalen schülerspezifischen Gewicht und dem schülerspezifischen Adjustierungsfaktor für nicht teilnehmende Schüler (OECD, 2014, S. 138).

3.3 Das Multi-Matrix-Design

Der Test von PISA 2012 soll latente Eigenschaften der zu prüfenden Personen erfassen. Unter latenten Personeneigenschaften versteht man Eigenschaften, die nicht direkt beobachtbar oder messbar sind. In diesem Fall handelt es sich um die Fertigkeiten der Schüler in den drei Kompetenzbereichen, aber auch deren Einstellungen und Überzeugungen. Hierbei resultiert bei der Studie ein Interessenskonflikt. Einerseits möchte man die latenten Eigenschaften möglichst umfassend prüfen, was ein breites Spektrum an Testaufgaben (Items) erfordert. Andererseits muss die Testzeit auf ein vernünftiges Maß begrenzt werden, schon alleine um die Teilnehmer nicht zu überfordern. Um diese Anforderungen in Einklang zu bringen wird ein sogenanntes Balanced Incomplete Block Design herangezogen (für Details hierzu siehe van der Linden, Veldkamp & Carlson, 2004). Dieses Design wurde entwickelt, um eine randomisierte, balancierte allerdings unvollständige Versuchsplanung durchzuführen. Hierbei müssen nicht alle Schüler sämtliche für den Test entwickelten Fragen beantworten, sondern lediglich eine bestimmte Auswahl aus dem verfügbaren Pool von Aufgaben. Es wird somit nicht jeder Schüler hinsichtlich aller zu untersuchender, latenter Eigenschaften geprüft. Es werden mehrere Testhefte erstellt, die eine unterschiedliche Auswahl von Aufgabengruppen in den einzelnen Testheften enthalten. Die nach dieser Art gestalteten Messinstrumente werden als Multi-Matrix-Design bezeichnet. Das Multi-Matrix-Design wurde entwickelt, um Daten mit unvollständiger, balancierter Struktur und einer großen Anzahl von Items zur Bestimmung von Populationsschätzwerten auf Basis der Erhebung großer Stichproben zu bestimmen (Heine et al., 2013, S.323f.).

Im Detail wurden die einzelnen Items zu den Bereichen Lesen, Mathematik und Naturwissenschaften um einen Aufgabenstamm (Testlet oder Unit) gruppiert. Diese Testlets oder Units enthalten Textelemente, Grafiken, Tabellen oder Kombinationen hieraus, wobei die Anzahl der Items pro Unit zwischen einer und sieben Aufgaben variieren konnte. Somit definieren die Units den geprüften Stoff bzw. die Themen und die Items die dazugehörigen spezifischen Fragen. Insgesamt wurden 87 Units erstellt, was 207 einzelnen Items entspricht (Heine, et al., 2013, S. 323f). Folgende Grafik soll dies nochmal verdeutlichen.

	Units gesamt	Items gesamt	Link-Units	Link-Items
Mathematik	56	110	25	36
Lesen	13	44	13	44
Naturwissenschaften	18	53	18	53
Gesamt	87	207	56	133

Fig. 4: Anzahl der an PISA 2012 eingesetzten Units und Items (Heine et al., 2013, S. 324)

Hierbei beschreiben die Link-Units und Link-Items die Themen und Aufgaben, die aus früheren PISA-Studien übernommen worden sind. In den Kompetenzbereichen Lesen und Naturwissenschaften wurden ausschließlich Link-Units verwendet. Hingegen sind für den Kompetenzbereich Mathematik mehr als 50% der Aufgaben neu erstellt worden (Heine et al., 2013, S. 324). In Grafik 4 wird zudem ersichtlich, dass bei PISA 2012 der Schwerpunkt auf Mathematik gelegt wurde.

3.4 Modellierung und Schätzung der latenten Personeneigenschaft

Das folgende Kapitel beschreibt Modelle um die latenten Personeneigenschaften zu schätzen. Da das Thema PISA-Studie generell ein sehr breites inhaltliches Spektrum umfasst und eine detaillierte Beschreibung der Modelle und besonders deren Schätzung den Rahmen dieser Seminararbeit sprengen würde, beschränkt sich dieses Kapitel auf die Grundgedanken dieser Modelle. Für Details sei auch auf die Seminararbeiten "Rasch-Modelle und Verallgemeinerung" sowie "Faktorenanalyse" verwiesen.
Das Multi-Matrix-Design bietet den Vorteil, dass eine breite Basis von Eigenschaften bei einem gleichzeitig begrenzten zeitlichen und inhaltlichen Umfang erreicht wird. Dies geschieht dadurch, indem nicht alle Schüler dieselben Aufgaben zur Bearbeitung bekommen. Hieraus erwächst allerdings auch direkt ein Nachteil für die folgende Testauswertung. Gegeben dass die Schüler nicht die gleichen Fragen beantworten und der Schwierigkeitsgrad zwischen den Testheften variieren kann, können die Leistungen der Schüler nicht direkt miteinander verglichen werden.
Um dennoch Aussagen über die latenten Eigenschaften der Schüler, besonders deren Fertigkeiten in den Bereichen Mathematik, Lesen und Naturwissenschaften, ermitteln zu können, werden das Rasch-Modell (Rasch, 1960) sowie dessen Erweiterung, das Partial-Credit-Modell (Masters, 1982), herangezogen. Beide beruhen auf der sogenannten Item-Response-Theory (IRT) (vgl. Fischer & Molenaar, 1995; Rost, 2004).

Das Rasch-Modell

Das Rasch-Modell stellt einen Zusammenhang zwischen den Fähigkeiten eines Schülers und der Aufgabenschwierigkeit her. Beides sind latente Eigenschaften für welche das Rasch-Modell Parameterschätzer liefert. Dabei beschreibt der Personenparameter θ die Fähigkeit der getesteten Person. Der sogenannte Itemparameter σ bezeichnet zudem den Schwierigkeitsgrad der Aufgabe. Das Rasch-Modell stellt eine plausible Beziehung zwischen den beiden Parametern her. Es modelliert die Wahrscheinlichkeit zur richtigen Lösung einer Aufgabe in Abhängigkeit der Abweichung dieser Para-meter voneinander. Dieser Modellierung liegt die Annahme zu Grunde, dass ein Schüler eine Aufgabe umso wahrscheinlicher richtig löst, je einfacher die Aufgabe ist (je niedriger σ) bzw. je höher dessen Fähigkeiten sind (hohes θ) (Heine et al., 2013). Die Beziehung der Parameter fließt als logistische Funktion in das Modell mit ein. Zur grafischen Darstellung verwendet man die Item-Characteristic Curve (ICC), welche wie folgt interpretiert werden kann: Mit zunehmender positiver Differenz zwischen dem Personenparameter und dem Itemparameter nimmt die Wahrscheinlichkeit der Lösung einer Aufgabe oberhalb der 50% Lösungswahrscheinlichkeit zu. Dies bedeutet, dass die Lösungswahrscheinlichkeit des betreffenden Items größer als 50% ist, wenn der Personenparameter größer ist, als der Itemparameter (Heine et al., 2013).

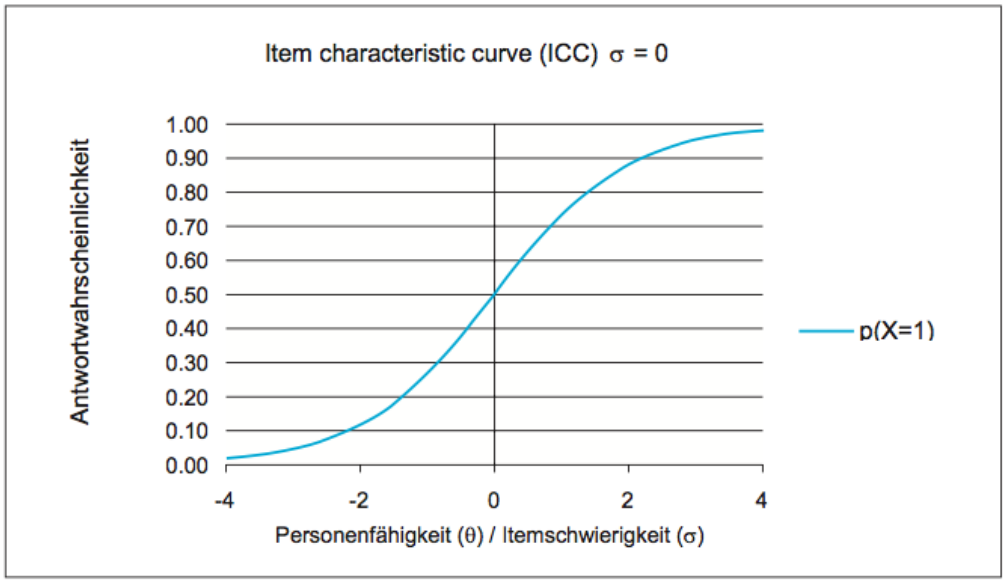

Fig. 5: Item-Characteristik Curve (ICC) (Heine et al., 2013, S. 327)

Formal:

σ : Itemparameter
θ : Personenparameter
$p(X)$: Lösungswahrscheinlichkeit
v : Person
i : Item

$$p(X_{vi}) = \frac{\exp(X_{vi}(\theta_v - \sigma_i))}{1 + \exp(\theta_v - \sigma_i)}, X \in 0, 1 \tag{1}$$

Hierbei bedeutet $X_{vi} = 0$, dass von einer Person v auf die Frage (Item) i eine falsche Antwort gegeben wurde und $X_{vi} = 1$, dass die gegebene Antwort richtig war. Da das *Rasch-Modell* keine teilweise richtigen Antworten zulässt, wird hierfür auf die Erweiterung des *Rasch-Modells* zurückgegriffen, das sogenannte *Partial-Credit-Modell* (Heine et al., 2013).

Das Partial-Credit-Modell

Das Partial-Credit-Modell lässt es zu mehrstufige Antwortformate mit zu berücksichtigen. So werden auch teilweise richtig gegebene Antworten von Schülern in der Testauswertung im Modell eingebettet. Um dies zu ermöglichen, wird der Itemparameter in einzelne Schwellenparameter zerlegt, die ein mehrstufiges, ordinales Antwortformat der Schüler zulassen (Heine et al., 2013).

Formal:

σ : Itemparameter
θ : Personenparameter
s : Schwellenwerte
σ_{ix} : realisierter Schwellenparameter
σ_{is} : möglicher Schwellenparameter

$$p(X_{vi} = x) = \frac{\exp((x\theta_v)) - \sigma_{ix}}{\sum_{s=0}^{m} \exp((s\theta_v) - \sigma_{is})}, x \in 0, 1, ...m \tag{2}$$

Die insgesamt m+1 Kategorien müssen aufsteigend von 0 bis m kodiert werden. Das Partial-Credit-Modell bestimmt somit die Wahrscheinlichkeit, dass ein Schüler v an der Schule i eine Antwort der Kategorie $x \in 0, ..., m$ gegeben hat. Dabei können die Antworten der Schüler als richtig, teilweise richtig oder falsch im Modell berücksichtigt werden (Heine et al., 2013, S. 328).

Die Modellparameter θ und σ werden mit der Maximum-Likelihood Methode bestimmt. Hierbei werden diese Parameter derart gewählt, dass die Wahrscheinlichkeit, die in das Modell eingeflossenen Testergebnisse zu beobachten, maximiert wird. Hierbei wird auf verschiedene Formen der Maximum-Likelihood-Schätzung wie der Joint-Maximum-Likelihood-Schätzung, der Conditional-Maximum-Likelihood-Schätzung oder der Marginal-Maximum-Likelihood Schätzung zurückgegriffen (Heine et al., 2013, S. 328ff.).

4 Skalierung und Einstufung der geschätzten latenten Personeneigenschaften

Nachdem die latenten Personeneigenschaften auf Basis der schriftlichen Testergebnisse mit Hilfe des Rasch-Modells bzw. des Partial-Credit-Modells geschätzt wurden, gilt es, diese verständlich und leicht interpretierbar zu gestalten. Hierbei werden die Schätzer für die latenten Eigenschaften für die Bereiche Mathematik, Lesen und Naturwissenschaften derart skaliert, dass diese jeweils um einen Mittelwert von 500 mit einer Standardabweichung von 100 schwanken (vgl. Heine et al., 2013, S. 339f). Um die Ergebnisse weiter besser interpretieren zu können, werden die Schüler zudem in den einzelnen Bereichen in Kompetenzstufen eingeteilt. Dies geschieht auf Basis des Grundgedankens des Rasch-Modells. Hierbei wird ein Schüler einer bestimmten Leistungsstufe zugeteilt, wenn er Aufgaben vom Schwierigkeitsgrad dieser Stufe mit einer Wahrscheinlichkeit von mehr als 50 Prozent, jedoch weniger als 70 Prozent lösen kann. Dabei wurde der Schwierigkeitsgrad einer Aufgabe ebenfalls zuvor durch das Rasch-Modell bzw. durch das Partial-Credit-Modell geschätzt. Für die einzelnen Kompetenzbereiche ergeben sich daraus folgende Leistungsgrenzen in Bezug auf die erreichte Punktezahl:

	Kompetenzbereich		
Kompetenzstufe	Mathematik	Lesen	Naturwissenschaften
6	669	698	708
5	607	626	633
4	545	553	559
3	482	480	484
2	420	407	409
1	358	1a: 335	335
		1b: 262	

Fig. 6: Cut-Off-Werte der Kompetenzstufen (Heine et al., 2013, S. 341)

Um diese Leistungsgrenzen anhand der skalierten Punktezahl anhand konkreter Kriterien zu veranschaulichen, zeigen die Abbildungen 11 - 13 im Anhang für jeden einzelnen Bereich die mit den Stufen jeweils einhergehenden Anforderungen und Fertigkeiten der Schüler.

5 Ergebnisse

Dieses Kapitel gibt einen Überblick über die wichtigsten Ergebnisse der PISA-Studie 2012. Hierbei werden besonders die im Mittel erreichten Punkte in den einzelnen Kompetenzbereichen betrachtet. Ein besonderes Augenmerk liegt dabei auf der relativen Leistung der Teilnehmerstaaten zueinander, weshalb die Ergebnisse in die folgenden Kategorien eingeteilt werden:

- statistisch signifikant **über** dem OECD-Durchschnittswert
- statistisch signifikant **unter** dem OECD-Durchschnittswert
- oder **nicht** vom OECD-Durchschnittswert **zu unterscheiden**

Neben dem im Mittel erreichten Punktewert und dem Abschneiden im internationalen Vergleich ist es zudem von großer Bedeutung die Standardabweichung der Leistungen der Schüler innerhalb eines Landes zu beachten (OECD, 2014). Denn diese gibt Auskunft darüber, wie groß das Leistungsspektrum der Schüler eines Landes sowie eventuell darüber wie stark das Bildungssystem innerhalb eines Landes variieren kann. Hierbei erfolgt die Betrachtung der Ergebnisse in den nachfolgenden Abschnitten getrennt nach Kompetenzbereichen.

5.1 Ergebnisse für den Kompetenzbereich Mathematik

Bei der PISA-Studie 2012 betrug im Kompetenzbereich Mathematik der Durchschnitt über alle Teilnehmerstaaten hinweg 494 Punkte. Die Abnahme des Durchschnitts gegenüber PISA 2003 (OECD-Durchschnitt auf 500 normiert) ist nicht unbedingt auf schlechtere Ergebnisse zurückzuführen, sondern kann auf die Erweiterung von vier Staaten (Chile, Estland, Israel und Slowenien) zurückgeführt werden, denn zwei dieser Staaten liegen deutlich unter dem OECD-Durchschnitt (Sälzer et al., 2013).

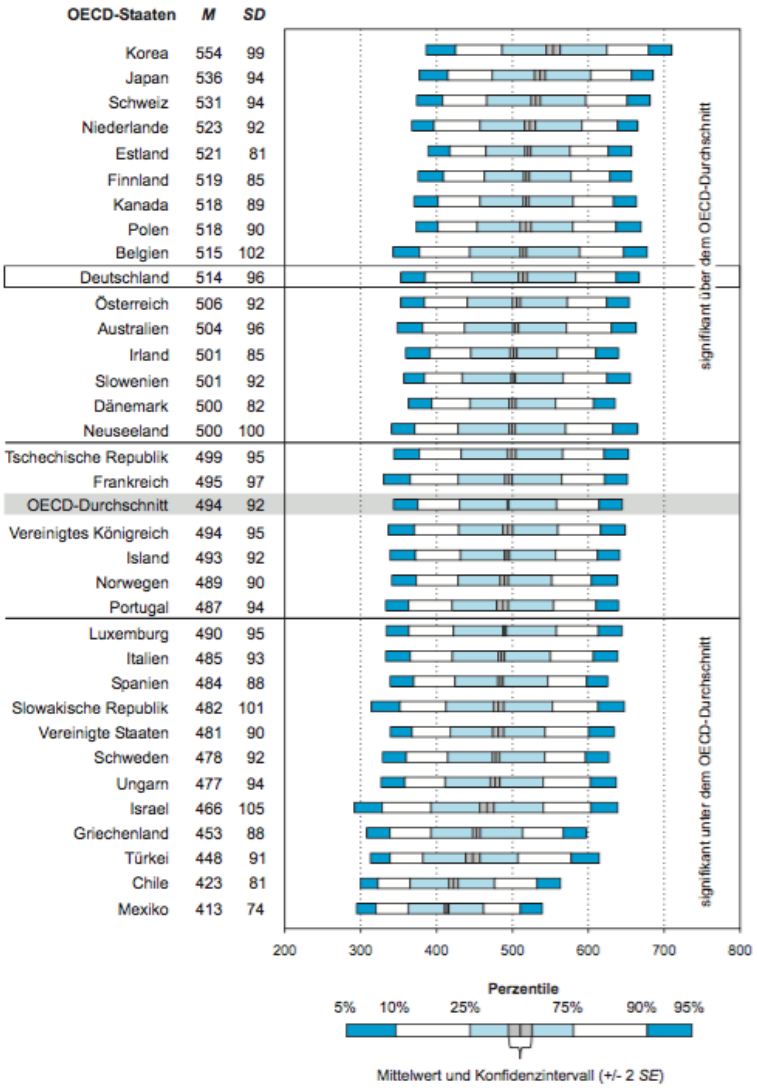

Fig. 7: Perzentilbänder der OECD-Staaten in der mathematischen Kompetenz (Sälzer et al., 2013, S. 71)

Mittelwerte im Vergleich

Obenstehende Grafik veranschaulicht die Mittelwerte der OECD-Staaten und deren Perzentilbänder. In grau sind die Mittelwerte und die beiden Konfidenzintervalle ersichtlich. Das Perzentilband eines jeden Staates zeigt auf, wo sich die 5, 10, 25, 75, 90 und 95% aller Schüler im jeweiligen Staat befinden. Diese Darstellungsweise hebt vor wie groß die Streuung der Kompetenz eines Landes ist. Bei 16 Staaten ist der Mittelwert statistisch signifikant über dem OECD-Durchschnitt von 494 Punkten ($p < 0.05$). Das Kompetenzniveau dieser Staaten reicht von 500 Punkten in Neuseeland bis hin zu 554 Punkten in Korea. Deutschland liegt mit einem Mittelwert von 514 Punkten statistisch signifikant über dem OECD-Durchschnitt und einer Standardabweichung von 96 Punkten im Bereich des OECD-Durchschnitts. Sechs Staaten unterscheiden sich nicht signifikant vom OECD-Durchschnitt. Mit 490 Punkten im Mittel liegt Luxemburg statistisch signifikant unter dem OECD-Durchschnitt. Mexiko schneidet mit einem Mittelwert von 413 Punkten am schlechtesten ab (Sälzer et al., 2013, S. 70ff).

Shanghai (China) belegte mit einer mittleren Punktzahl von 613 Punkten den ersten Platz im Bereich Mathematik. Direkt darauf folgen Singapur (mit 573 Punkten) und Hongkong (mit 561 Punkten). Deutschland belegte Platz 16 mit einer mittleren Punktzahl von 514 Punkten. Bei 25 der 64 teilnehmenden Staaten, für die Trenddaten in dem Zeitraum von 2003 bis 2012 vorlagen, war eine Verbesserung im Kompetenzbereich Mathematik zu verzeichnen. 13% der Schüler sind hierbei als "besonders leistungsstark" einzuordnen, was bedeutet, dass im Durchschnitt aller OECD-Staaten jene Schüler die Kompetenzstufe 5 oder 6 aufweisen. Hierbei weist Shanghai den größten Anteil an Schülern mit der höchsten Kompetenzstufe auf. Hier erreichten 55% der PISA-Teilnehmer die Kompetenzstufe 5 oder höher. Singapur verzeichnet einen Anteil von 40% der Schüler in diesen Kompetenzstufen. Hingegen gelten in Deutschland nur $17,5\%$ der Schüler als "besonders leistungsstark" wobei hier die Jungen mit 20% besser abschneiden als die Mädchen mit 15% (OECD, 2013).
Insgesamt ist es 23% der Schüler in den OECD-Ländern nicht gelungen die zweite Kompetenzstufe im Bereich Mathematik zu erzielen. In Deutschland lag der Anteil der "besonders leistungsschwachen" Schüler, d.h. Schüler die die Kompetenzstufe 2 nicht erreichen konnten, im Kompetenzbereich Mathematik bei 17,7%. Bei den geschlechtsspezifischen Unterschieden unter den "besonders leistungsschwachen" Schülern sind die Mädchen mit 19% stärker vertreten als die Jungen mit 17% (OECD, 2013).

Zwischen 2003 und 2012 war in Deutschland dennoch eine signifikante Verbesserung der Mathematikleistungen zu erkennen. Die annualisierte Veränderung beträgt 1,4 Punkte (OECD, 2013).
Allgemein waren die Leistungen in 37 der 65 Teilnehmerstaaten im Jahr 2012 die männlichen Schüler besser als die der weiblichen Schüler. In nur 5 Länder schnitten die Mädchen besser ab im Kompetenzbereich Mathematik als die Jungen. Die Jungen in Deutschland erreichten durchschnittlich 14 Punkte mehr als die Mädchen. Im OECD-Mittel erzielten die männlichen Schüler 11 Punkte mehr als die weiblichen (OECD, 2013).
Das schlechteste Ergebnis und somit den letzten Rang erreichte Peru mit 368 Punkten und lag somit deutlich unter dem OECD-Durchschnitt (OECD, 2013). Eine Tabelle mit allen Ergebnissen im Überblick findet sich im Anhang (Tabelle 10).

Streuungen im Kompetenzbereich Mathematik

Für Deutschland ergab sich in PISA 2012 eine Standardabweichung von 96 Punkten. Somit liegt Deutschland über dem OECD-Durchschnitt von 92 Punkten. In Korea und Belgien lassen sich große Streuungen erkennen, wobei Belgien eine Standardabweichung von 102 Punkten aufweist und Korea eine Standardabweichung von 99 Punkten. Geringe Streuungen hingegen weisen Estland ($SD = 81\ Punkte$), Dänemark ($SD = 82\ Punkte$), Finnland ($SD = 85\ Punkte$) und Irland ($SD = 85\ Punkte$) auf (Sälzer et al., 2013, S. 73). Weiter ist erkennbar, dass die Streuungen in allen Staaten sehr groß ist, sich die Perzentilbänder aller Staaten aber auch stark überlappen. Die größte Streuung weist Israel auf, während Mexiko die geringste Streuung hat (Sälzer et al., 2013, S. 70ff).

5.2 Ergebnisse für den Kompetenzbereich Lesen

Der OECD-Durchschnitt in dem Kompetenzbereich Lesen lag in PISA 2012 bei 496 Punkten. Die Standardabweichung lag bei 94 Punkten. Shanghai belegte auch hier wieder den ersten Rang mit 570 erreichten Punkten. Danach folgten Hongkong (545 Punkte), Singapur (542 Punkte) und Japan (538 Punkte). Deutschland erreichte in der Kompetenz Lesen 508 Punkte und lag damit knapp über dem OECD-Durchschnitt. Auch hier war eine signifikante annualisierte Veränderung von 1,8 Punkten zu vermerken. Einer Verbesserung der Lesekompetenz war in 32 der 64 teilnehmenden Ländern für die Vergleichsdaten vorlagen zu erkennen. Den letzten Platz belegte Peru mit 384 erzielten Punkten. Dennoch lies sich hier eine signifikante

annualisierte Verbesserung von 5,2 Punkten verzeichnen (OECD, 2013).
Nur 8% der Schüler erreichten in dieser Kompetenz die Kompetenzstufe 5 oder 6. Den größten Anteil der "besonders leistungsstarken" Schüler verzeichnete Shanghai mit 25%. Der Anteil der Schüler in den Kompetenzstufen 5 oder 6 verbesserte sich in den Jahren 2000 und 2012 in den Ländern Albanien, Israel und Polen, wobei sich der Anteil der "besonders leistugsschwachen" Schüler gleichzeitig verringerte. In den Jahren 2000 und 2012 verbesserte sich die Leistung der weiblichen Schüler in dieser Kompetenz in elf Ländern (OECD, 2013).

Mittelwerte im Vergleich

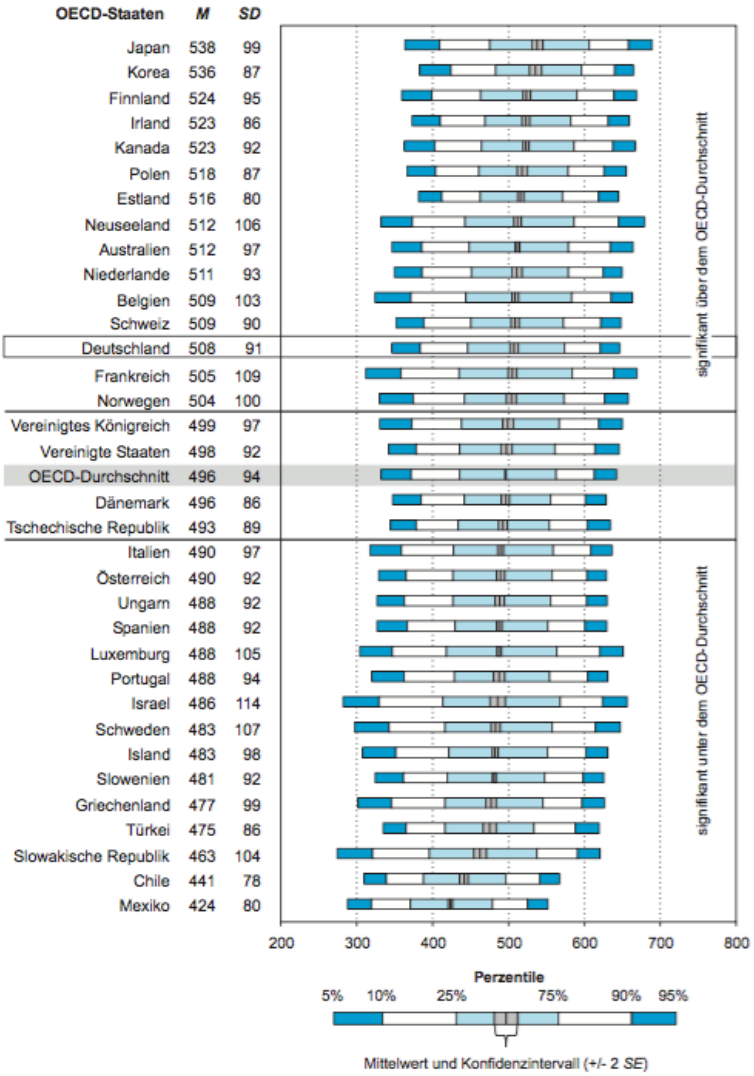

Fig. 8: Perzentilbänder der Lesekompetenz in den OECD-Staaten (Hohn et al., 2013, S. 219f)

Grafik 8 zeigt die Perzentilbänder der OECD-Staaten für den Kompetenzbereich Lesen. 15 Staaten lagen in PISA 2012 signifikant über dem OECD-Durchschnitt, 15 Staaten lagen unter dem OECD-Durchschnitt und 4 Staaten weichen nicht signifikant davon ab. Deutschland liegt mit 508 Punkten erstmals signifikant über dem Durchschnitt, wodurch im internationalen Vergleich von einer positiven Entwicklung Deutschlands gesprochen werden kann (Hohn et al., 2013, S. 228.).
Die letzten Ränge des Kompetenzbereichs Lesen unter den OECD-Staaten belegen Mexiko mit 424 Punkten und Chile mit 441 Punkten. Eine ausführliche Tabelle der Ränge der OECD-Staaten und Partnerstaaten ist im Anhang zu finden (Tabelle 10).

Streuungen im Kompetenzbereich Lesen

Die Standardabweichung gilt als Maß der Streuung und ist neben dem Mittelwert eine wichtige Maßzahl um Unterschiede innerhalb der einzelnen Staaten zu erfassen. Hohe Werte der Streuung weisen auf große Unterschiede innerhalb eines Staates hin, wohingegen kleine Werte geringere Unterschiede zwischen den leistungsstärksten und den leistungsschwächsten Schülern eines Landes bedeuten. In PISA 2012 betrug die durchschnittliche Streuung 94 Punkte (Hohn et al., 2013, S. 228).
Besonders hohe Streuungen unter den OECD-Staaten weisen Frankreich ($SD = 109$), Schweden($SD = 107$) und Luxemburg ($SD = 105$) auf. Auffällig sind die hohen Werte bei Japan ($SD = 99$) und Korea ($SD = 87$), die die obersten Ränge in PISA 2012 belegen (Hohn et al., 2013, S. 228).
Deutschland wies mit 91 Punkten eine signifikant vom Durchschnitt kleinere Streuung auf. Die positive Entwicklung wird durch die Verringerung der Streuung aufgezeigt.
Grafik 8 zeigt zudem die einzelnen Konfidenzintervalle der OECD-Staaten zwischen den 5% der leistungsstärksten Schüler und den 5% der leistungsschwächsten Schüler sowie dessen Mittelwert.
Insgesamt könnten in diesem Kompetenzbereich ca. 14% der Schüler in Deutschland Kompetenzstufe 2 nicht erreichen. Das OECD-Mittel der Schüler die diese Kompetenzstufe verfehlten liegt bei 18%. Jedoch können insgesamt etwa 9% der Schülerinnen und Schüler Kompetenzstufe 5 oder höher erreichen.
In der Lesekompetenz hatten die Mädchen im Durchschnitt 44 Punkte mehr erreicht als die Jungen. Der OECD-Durchschnitt der leistungsspezifischen Unterschiede liegt bei 38 Punkten zugunsten der weiblichen Schülerinnen. Des Weiteren ist der Anteil der Mädchen in den Kompetenzstufen 5 oder höher bei 13%, während der Anteil der Jungen hier bei 5% liegt.

5.3 Ergebnisse für den Kompetenzbereich Naturwissenschaften

Im Bereich Naturwissenschaften war der Durchschnitt über alle OECD-Staaten hinweg 501 Punkte. Die fünf besten Länder im Kompetenzbereich Naturwissenschaften waren Shanghai (580 Punkte), Hongkong (555 Punkte), Singapur (551 Punkte), Japan (547 Punkte) und Finnland (545 Punkte). Deutschland erreichte mit 524 Punkten Rang elf und liegt signifikant oberhalb des OECD-Durchschnitts. Die annualisierte Veränderung Deutschlands lag bei 1,4 Punkten und war somit höher als der OECD-Durchschnitt von 0.5 Punkten (OECD, 2013).
In Italien, Polen und Katar verbesserte sich der Anteil der Schüler in den Kompetenzstufen 5 und 6 in den Jahren zwischen 2006 und 2012. Der Anteil der Schüler in den Kompetenzstufen unter 2 verringerte sich gleichzeitig im Kompetenzbereich Naturwissenschaften. Insgesamt konnten sich 8% der Schüler der OECD-Teilnehmer zu den "besonders leistungsstarken" Schüler zählen (OECD, 2013).

Mittelwerte im Vergleich

16 Staaten erreichten in PISA 2012 mittlere Kompetenzwerte, die den OECD-Durchschnitt signifikant übertreffen. Die vier besten OECD-Staaten sind Japan (547 Punkte), Finnland (545 Punkte), Estland (541 Punkte) und Korea (538 Punkten). Bis zu einer halben Standardabweichung liegen diese Staaten mit den Kompetenzwerten über dem OECD-Durchschnitt (Schiepe-Tiska et al., 2013).
Polen (526 Punkte) und Kanada (525 Punkte) unterscheiden sich nicht signifikant von Deutschland. Mexiko erreichte den niedrigsten mittleren Kompetenzwert mit 415 Punkten (Schiepe-Tiska et al., 2013).

Die Streuung naturwissenschaftlicher Kompetenz

Die Homogenität der einzelnen Staaten lässt sich an der Standardabweichung festlegen. Der OECD-Durchschnitt der Streuung in den OECD-Staaten beträgt 93 Punkte. Deutschland erreichte eine Standardabweichung von 95 Punkten und ist im Bereich des OECD-Durchschnitts. Besonders heterogene Verteilungen sind in Israel ($SD = 108$), Neuseeland ($SD = 105$) und Luxemburg ($SD = 103$) gemessen worden. Die Standardabweichungen der naturwissenschaftliche Kompetenz in Australien

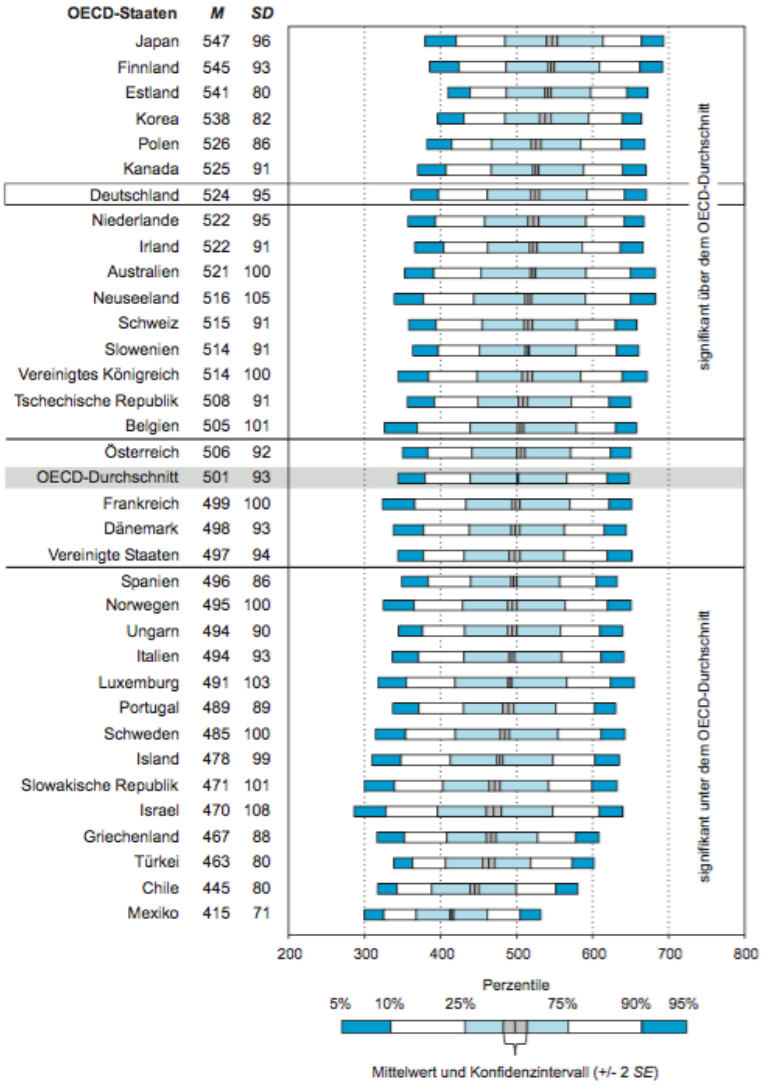

Fig. 9: Perzentilbänder der naturwissenschaftlichen Kompetenz (Schiepe-Tiska et al., 2013, S. 199)

($SD = 100$), Belgien ($SD = 101$), dem Vereinten Königreich ($SD = 100$), Frankreich ($SD = 100$), Norwegen ($SD = 100$), Schweden ($SD = 100$), der Slowakischen Republik ($SD = 101$) und Island ($SD = 99$) streuen signifikant stärker als im OECD-Mittel. In Estland ($SD = 80$), Korea ($SD = 82$) und Polen ($SD = 86$) hingegen zeigt sich eine besonders homogene Verteilung. Die Staaten der letzten Ränge weisen eine geringe Streuung auf mit Werten von ($SD = 80$) Punkten für die Türkei & Chile und ($SD = 71$) Punkten für Mexiko (Schiepe-Tiska et al., 2013).

In Deutschland konnten etwa 12% der Schüler die Mindestanforderungen nicht erfüllen. Durchschnittlich verzeichnete die OECD, dass in allen Teilnehmerstaaten 18% den Leistungsanforderungen an Kompetenzstufe II nicht gerecht wurden. Hierbei war

der Anteil der männlichen Schüler bei 13% und der der weiblichen Schüler bei 11%. Ebenfalls ca. 12% der Schüler in Deutschland konnten als "besonders leistungsstark" eingestuft werden. Ein Unterschied zwischen den Geschlechtern konnte hier nicht festgestellt werden.

Eine Grafik, die alle Ergebnisse im Überblick zeigt, wobei die Länder und Volkswirtschaften in absteigender Reihenfolge nach den Durchschnittsergebnissen im Bereich Mathematik in PISA 2012 angeordnet sind, ist im Anhang zu finden.

6 Ausblick

Diese Arbeit gibt einen Überblick über die grundlegende Durchführung der PISA-Studie des Jahres 2012 sowie die hierbei eingesetzte Methodik. Aufgrund des enormen Umfangs und der teilweise hohen Komplexität der Methodik kann und soll diese Arbeit nicht eine umfassende und tiefgreifende Erläuterung dieser geben. Vielmehr wird die Methodik derart zusammengefasst, dass das grundlegende methodische Vorgehen der Studie verständlich wird.

Im Zuge der breiten öffentlichen Debatte über die Ergebnisse und Konsequenzen der PISA-Studie, wurde an dieser auch oftmals Kritik geübt. Hierbei geben jedoch Sälzer & Prenzel zu bedenken, dass es innerhalb der Fachliteratur keine fundamentale Kritik hinsichtlich des methodischen Vorgehens geäußert wird. Sie folgern, dass künftiger Handlungsbedarf weniger in der Verbesserung der Methodik liegt, sondern vielmehr darin, dieses komplexe Thema der Öffentlichkeit besser verständlich zu machen, um Fehlinterpretationen entgegen zu wirken (Sälzer & Prenzel, 2013, S. 20).

Nichtsdestotrotz ist die PISA-Studie nicht frei von Kritik. Genannt werden kann hier unter anderem, dass Probleme bei der Übersetzung der Testaufgaben wie etwa Übersetzungsfehler oder aber deutliche Abweichungen in der Textlänge zwischen einzelnen Sprachen auftreten können. Auch unterschiedliche kulturelle Gewohnheiten können die Vergleichbarkeit erschweren. Wuttke (2006) merkt hierzu an, dass die Aufgaben stark vom anglo-amerikanischen Bildungssystem geprägt sind und somit Schüler anderer Kulturkreise mit dem Prüfungsmuster weniger vertraut sind. Generell ist ebenso denkbar, dass kulturelle Unterscheide wie auch abweichende Bildungssysteme dazu führen können, dass es starke Abweichungen in der Motivation der Schüler gibt. Sollten sich die Schüler eines Landes beim Test mehr anstrengen als andere, könnte dies dazu führen, dass dies die geschätzte Kompetenz verzerrt.

Somit bleibt festzustellen, dass es noch weiterer Anstrengungen bedarf, um die Interpretierbarkeit und Vergleichbarkeit der PISA-Studie zu erhöhen, auch wenn diese sich weniger auf das statistische Vorgehen beziehen, sondern vielmehr auf sprachliche und interkulturelle Differenzen zwischen den Teilnehmerstaaten.

7 Anhang

Länder/Volkswirtschaften, deren Durchschnittsergebnis/Anteil besonders leistungsstarker Schüler über dem OECD-Durchschnitt liegt
Länder/Volkswirtschaften, deren Anteil besonders leistungsschwacher Schüler unter dem OECD-Durchschnitt liegt
Länder/Volkswirtschaften, deren Durchschnittsergebnis/Anteil besonders leistungsschwacher bzw. leistungsstarker Schüler nicht statistisch signifikant vom OECD-Durchschnitt abweicht
Länder/Volkswirtschaften, deren Durchschnittsergebnis/Anteil besonders leistungsstarker Schüler unter dem OECD-Durchschnitt liegt
Länder/Volkswirtschaften, deren Anteil besonders leistungsschwacher Schüler über dem OECD-Durchschnitt liegt

	Mathematik				Lesekompetenz		Naturwissenschaften	
	Mittelwert PISA 2012	Anteil besonders leistungsschwacher Schüler (unter Stufe 2)	Anteil besonders leistungsstarker Schüler (Stufe 5 und 6)	Annualisierte Veränderung in Punkten	Mittelwert PISA 2012	Annualisierte Veränderung in Punkten	Mittelwert PISA 2012	Annualisierte Veränderung in Punkten
OECD-Durchschnitt	494	23.0	12.6	**-0.3**	496	**0.3**	501	**0.5**
Shanghai (China)	613	3.8	55.4	**4.2**	570	**4.6**	580	1.8
Singapur	573	8.3	40.0	**3.8**	542	**5.4**	551	**3.3**
Hongkong (China)	561	8.5	33.7	**1.3**	545	**2.3**	555	**2.1**
Chinesisch Taipeh	560	12.8	37.2	1.7	523	**4.5**	523	-1.5
Korea	554	9.1	30.9	1.1	536	**0.9**	538	**2.6**
Macau (China)	538	10.8	24.3	**1.0**	509	**0.8**	521	**1.6**
Japan	536	11.1	23.7	0.4	538	**1.5**	547	**2.6**
Liechtenstein	535	14.1	24.8	0.3	516	**1.3**	525	0.4
Schweiz	531	12.4	21.4	0.6	509	**1.0**	515	0.6
Niederlande	523	14.8	19.3	**-1.6**	511	-0.1	522	-0.5
Estland	521	10.5	14.6	0.9	516	**2.4**	541	1.5
Finnland	519	12.3	15.3	**-2.8**	524	**-1.7**	545	**-3.0**
Kanada	518	13.8	16.4	**-1.4**	523	**-0.9**	525	**-1.5**
Polen	518	14.4	16.7	**2.6**	518	**2.8**	526	**4.6**
Belgien	515	19.0	19.5	**-1.6**	509	0.1	505	-0.9
Deutschland	514	17.7	17.5	**1.4**	508	**1.8**	524	1.4
Vietnam	511	14.2	13.3	m	508	m	528	m
Österreich	506	18.7	14.3	0.0	490	-0.2	506	-0.8
Australien	504	19.7	14.8	**-2.2**	512	**-1.4**	521	-0.9
Irland	501	16.9	10.7	-0.6	523	**-0.9**	522	**2.3**
Slowenien	501	20.1	13.7	-0.6	481	**-2.2**	514	-0.8
Dänemark	500	16.8	10.0	**-1.8**	496	0.1	498	0.4
Neuseeland	500	22.6	15.0	**-2.5**	512	**-1.1**	516	**-2.5**
Tschech. Rep.	499	21.0	12.9	**-2.5**	493	-0.5	508	-1.0
Frankreich	495	22.4	12.9	**-1.5**	505	0.0	499	0.6
Ver. Königreich	494	21.8	11.8	-0.3	499	0.7	514	-0.1
Island	493	21.5	11.2	**-2.2**	483	**-1.3**	478	**-2.0**
Lettland	491	19.9	8.0	0.5	489	**1.9**	502	**2.0**
Luxemburg	490	24.3	11.2	-0.3	488	**0.7**	491	0.9
Norwegen	489	22.3	9.4	-0.3	504	0.1	495	1.3
Portugal	487	24.9	10.6	**2.8**	488	**1.6**	489	**2.5**
Italien	485	24.7	9.9	**2.7**	490	0.5	494	**3.0**
Spanien	484	23.6	8.0	0.1	488	-0.3	496	1.3
Russ. Föderation	482	24.0	7.8	1.1	475	**1.1**	486	1.0
Slowak. Rep.	482	27.5	11.0	**-1.4**	463	-0.1	471	**-2.7**
Ver. Staaten	481	25.8	8.8	0.3	498	-0.3	497	1.4
Litauen	479	26.0	8.1	-1.4	477	1.1	496	1.3
Schweden	478	27.1	8.0	**-3.3**	483	**-2.8**	485	**-3.1**
Ungarn	477	28.1	9.3	**-1.3**	488	**1.0**	494	-1.6
Kroatien	471	29.9	7.0	0.6	485	1.2	491	-0.3
Israel	466	33.5	9.4	**4.2**	486	**3.7**	470	**2.8**
Griechenland	453	35.7	3.9	**1.1**	477	0.5	467	-1.1
Serbien	449	38.9	4.6	**2.2**	446	**7.6**	445	1.5
Türkei	448	42.0	5.9	**3.2**	475	**4.1**	463	**6.4**
Rumänien	445	40.8	3.2	**4.9**	438	1.1	439	**3.4**
Zypern[1, 2]	440	42.0	3.7	m	449	m	438	m
Bulgarien	439	43.8	4.1	**4.2**	436	0.4	446	2.0
Ver. Arab. Emirate	434	46.3	3.5	m	442	m	448	m
Kasachstan	432	45.2	0.9	**9.0**	393	0.8	425	**8.1**
Thailand	427	49.7	2.6	1.0	441	**1.1**	444	**3.9**
Chile	423	51.5	1.6	**1.9**	441	**3.1**	445	1.1
Malaysia	421	51.8	1.3	**8.1**	398	**-7.8**	420	-1.4
Mexiko	413	54.7	0.6	**3.1**	424	**1.1**	415	0.9
Montenegro	410	56.6	1.0	**1.7**	422	**5.0**	410	-0.3
Uruguay	409	55.8	1.4	**-1.4**	411	**-1.8**	416	**-2.1**
Costa Rica	407	59.9	0.6	-1.2	441	-1.0	429	-0.6
Albanien	394	60.7	0.8	**5.6**	394	**4.1**	397	2.2
Brasilien	391	67.1	0.8	**4.1**	410	**1.2**	405	**2.3**
Argentinien	388	66.5	0.3	1.2	396	-1.6	406	2.4
Tunesien	388	67.7	0.8	**3.1**	404	**3.8**	398	**2.2**
Jordanien	386	68.6	0.6	0.2	399	-0.3	409	**-2.1**
Kolumbien	376	73.8	0.3	1.1	403	**3.0**	399	1.8
Katar	376	69.6	2.0	**9.2**	388	**12.0**	384	**5.4**
Indonesien	375	75.7	0.3	0.7	396	**2.3**	382	-1.9
Peru	368	74.6	0.6	1.0	384	**5.2**	373	1.3

Fig. 10: Ergebnis der PISA-Erhebung 2012 in den drei Kompetenzbereichen (OECD, 2013, S. 5)

Kompetenzstufe	Wozu die Schülerinnen und Schüler auf der jeweiligen Kompetenzstufe im Allgemeinen in der Lage sind
VI > 698 Punkte	Jugendliche auf dieser Stufe können Schlussfolgerungen, Vergleiche und Gegenüberstellungen detailgenau und präzise anstellen. Dabei entwickeln sie ein volles und detailliertes Verständnis eines oder mehrerer Texte und verbinden dabei unter Umständen gedanklich Informationen aus mehreren Texten miteinander. Hierbei kann auch die Auseinandersetzung mit ungewohnten Ideen gefordert sein, genauso wie der kompetente Umgang mit konkurrierenden Informationen und abstrakten Interpretationskategorien sowie hohe Präzision im Umgang mit zum Teil unauffälligen Textdetails.
V 626–698 Punkte	Jugendliche auf dieser Stufe können sowohl mehrere tief eingebettete Informationen finden, ordnen und herausfinden, welche davon jeweils relevant sind, als auch ausgehend von Fachwissen eine kritische Beurteilung oder Hypothese anstellen. Die Aufgaben dieser Stufe setzen in der Regel ein volles und detailliertes Verständnis von Texten voraus, deren Inhalt oder Form ungewohnt ist. Zudem muss mit Konzepten umgegangen werden können, die im Gegensatz zum Erwarteten stehen.
IV 553–626 Punkte	Aufgaben dieser Kompetenzstufe erfordern vom Leser/von der Leserin, linguistischen oder thematischen Verknüpfungen in einem Text über mehrere Abschnitte zu folgen, oftmals ohne Verfügbarkeit eindeutiger Kennzeichen im Text, um eingebettete Informationen zu finden, zu interpretieren und zu bewerten oder um psychologische oder philosophische Bedeutungen zu erschließen. Insgesamt muss ein genaues Verständnis langer oder komplexer Texte, deren Inhalt oder Form ungewohnt sein kann, unter Beweis gestellt werden.
III 480–553 Punkte	Aufgaben dieser Kompetenzstufe erfordern vom Leser/von der Leserin, vorhandenes Wissen über die Organisation und den Aufbau von Texten zu nutzen, implizite oder explizite logische Relationen (z. B. Ursache-Wirkungs-Beziehungen) über mehrere Sätze oder Textabschnitte zu erkennen, mit dem Ziel, Informationen im Text zu lokalisieren, zu interpretieren und zu bewerten. Einige Aufgaben verlangen vom Leser/von der Leserin, einen Zusammenhang zu begreifen oder die Bedeutung eines Wortes oder Satzes zu analysieren. Häufig sind die benötigten Informationen dabei nicht leicht sichtbar oder Passagen des Textes laufen eigenen Erwartungen zuwider.
II 407–480 Punkte	Jugendliche auf dieser Stufe können innerhalb eines Textabschnitts logischen und linguistischen Verknüpfungen folgen, mit dem Ziel, Informationen im Text zu lokalisieren oder zu interpretieren, im Text oder über Textabschnitte verteilte Informationen aufeinander beziehen, um die Absicht des Autors zu erschließen. Bei Aufgaben dieser Stufe müssen unter Umständen auf der Grundlage eines einzigen Textbestandteils Vergleiche und Gegenüberstellungen vorgenommen werden oder es müssen, ausgehend von eigenen Erfahrungen oder Standpunkten, Vergleiche angestellt oder Zusammenhänge zwischen dem Text und nicht im Text enthaltenen Informationen erkannt werden.
Ia 335–407 Punkte	Aufgaben dieser Kompetenzstufe erfordern vom Leser/von der Leserin, in einem Text zu einem vertrauten Thema eine oder mehrere unabhängige, explizit ausgedrückte Informationen zu lokalisieren, das Hauptthema oder die Absicht des Autors zu erkennen oder einen einfachen Zusammenhang zwischen den im Text enthaltenen Informationen und allgemeinem Alltagswissen herzustellen. Die erforderlichen Informationen sind in der Regel leicht sichtbar, und es sind nur wenige beziehungsweise keine konkurrierenden Informationen vorhanden. Der Leser wird explizit auf die entscheidenden Elemente in der Aufgabe und im Text hingewiesen.
Ib 262–335 Punkte	Jugendliche auf dieser Stufe können in einem kurzen, syntaktisch einfachen Text aus einem gewohnten Kontext, dessen Form vertraut ist (z. B. in einer einfachen Liste oder Erzählung), eine einzige, explizit ausgedrückte Information lokalisieren, die leicht sichtbar ist. Der Text enthält in der Regel Hilfestellungen für den Leser, wie Wiederholungen, Bilder oder bekannte Symbole. Es gibt kaum konkurrierende Informationen. Bei anderen Aufgaben müssen einfache Zusammenhänge zwischen benachbarten Informationsteilen hergestellt werden.

Fig. 11: Überblick über die Anforderungen pro Kompetenzstufe im Bereich Lesen (Hohn et al., 2013, S. 219f.)

Kompetenzstufe	Wozu die Schülerinnen und Schüler auf der jeweiligen Kompetenzstufe im Allgemeinen in der Lage sind
VI > 669 Punkte	Schülerinnen und Schüler auf dieser Stufe können Informationen, die sie aus der Untersuchung und Modellierung komplexer Problemsituationen erhalten, konzeptualisieren, verallgemeinern und auf neue Situationen anwenden. Sie können verschiedene Informationsquellen und Darstellungen miteinander verknüpfen und flexibel zwischen diesen hin und her wechseln. Schülerinnen und Schüler auf dieser Stufe besitzen die Fähigkeit zu anspruchsvollem mathematischem Denken und Argumentieren. Sie können dieses mathematische Verständnis und ihre Beherrschung symbolischer und formaler mathematischer Operationen und Beziehungen nutzen, um Ansätze und Strategien zum Umgang mit neuartigen Problemsituationen zu entwickeln. Schülerinnen und Schüler auf dieser Stufe können ihr Tun und ihre Überlegungen, die zu ihren Erkenntnissen, Interpretationen und Argumentationen geführt haben, präzise beschreiben und kommunizieren, einschließlich der Beurteilung von deren Angemessenheit für die jeweilige Ausgangssituation.
V 607–668 Punkte	Schülerinnen und Schüler auf dieser Stufe können Modelle für komplexe Situationen konzipieren und mit ihnen arbeiten, einschränkende Bedingungen identifizieren und Annahmen spezifizieren. Sie können im Zusammenhang mit diesen Modellen geeignete Strategien für die Lösung komplexer Probleme auswählen, sie miteinander vergleichen und bewerten. Schülerinnen und Schüler auf dieser Stufe können strategisch vorgehen, indem sie sich auf breit gefächerte, gut entwickelte Denk- und Argumentationsfähigkeiten, passende Darstellungen, symbolische und formale Beschreibungen und für diese Situationen relevante Einsichten stützen. Sie sind imstande, über ihr Tun zu reflektieren und ihre Interpretationen und Überlegungen zu formulieren und zu kommunizieren.
IV 545–606 Punkte	Schülerinnen und Schüler auf dieser Stufe können effektiv mit expliziten Modellen komplexer konkreter Situationen arbeiten, auch wenn sie einschränkende Bedingungen enthalten oder die Aufstellung von Annahmen erfordern. Sie können verschiedene Darstellungsformen, darunter auch symbolische, auswählen und zusammenführen, indem sie sie direkt zu Aspekten von Realsituationen in Beziehung setzen. Schülerinnen und Schüler auf dieser Stufe können in diesen Kontexten gut ausgebildete Fertigkeiten anwenden und mit einem gewissen mathematischen Verständnis flexibel argumentieren. Sie können Erklärungen und Begründungen für ihre Interpretationen, Argumentationen und Handlungen geben und sie anderen mitteilen.
III 483–544 Punkte	Schülerinnen und Schüler auf dieser Stufe können klar beschriebene Verfahren durchführen, auch solche, die sequenzielle Entscheidungen erfordern. Sie können einfache Problemlösungsstrategien auswählen und anwenden. Schülerinnen und Schüler auf dieser Stufe können Darstellungen interpretieren und nutzen, die aus verschiedenen Informationsquellen stammen, und hieraus unmittelbare Schlüsse ableiten. Sie können kurze Berichte zu ihren Interpretationen, Ergebnissen und Überlegungen geben.
II 421–482 Punkte	Schülerinnen und Schüler auf dieser Stufe können Situationen in Kontexten interpretieren und erkennen, die nicht mehr als direkte Schlussfolgerungen erfordern. Sie können relevante Informationen einer einzigen Quelle entnehmen und eine einzige Darstellungsform benutzen. Schülerinnen und Schüler auf dieser Stufe können elementare Algorithmen, Formeln, Verfahren oder Regeln anwenden. Sie sind zu direkten Schlussfolgerungen und wörtlichen Interpretationen der Ergebnisse imstande.
I 358–420 Punkte	Schülerinnen und Schüler auf dieser Stufe können auf Fragen zu vertrauten Kontexten antworten, bei denen alle relevanten Informationen gegeben und die Fragen klar definiert sind. Sie können Informationen identifizieren und Routineverfahren gemäß direkten Instruktionen in expliziten Situationen anwenden. Sie können Handlungen ausführen, die klar ersichtlich sind und sich unmittelbar aus den jeweiligen Situationen ergeben.
unter I < 358 Punkte	

Fig. 12: Stufen mathematischer Kompetenz (Sälzer et al., 2013, S. 61)

Kompetenzstufe	Wozu die Schülerinnen und Schüler auf der jeweiligen Kompetenzstufe im Allgemeinen in der Lage sind
VI > 707 Punkte	Auf Stufe VI können Schülerinnen und Schüler in konsistenter Weise naturwissenschaftliches Wissen und Wissen über die Naturwissenschaften in einer Vielzahl komplexer Lebenssituationen erkennen, erklären und anwenden. Sie können verschiedene Informationsquellen und Erklärungen zueinander in Beziehung setzen und die Beweise, die aus diesen Quellen folgen, nutzen, um Entscheidungen zu begründen. Sie demonstrieren ein weit entwickeltes naturwissenschaftliches und logisches Denkvermögen und sind bereit, ihr naturwissenschaftliches Verständnis einzusetzen, um Lösungen für unbekannte naturwissenschaftliche oder technologische Probleme zu finden. Schülerinnen und Schüler auf dieser Stufe können naturwissenschaftliches Wissen anwenden und Argumente entwickeln, um Empfehlungen auszusprechen beziehungsweise Entscheidungen zu treffen, die von persönlicher, sozialer oder globaler Bedeutung sind.
V 634–707 Punkte	Auf Stufe V können Schülerinnen und Schüler die Bedeutung der Naturwissenschaften in komplexen Lebenssituationen erkennen. Sie können sowohl ihr konzeptuelles Wissen als auch ihr Wissen über die Naturwissenschaften auf diese Situationen anwenden und naturwissenschaftliche Beweise vergleichen, auswählen und bewerten, um in angemessener Weise auf diese Situationen zu reagieren. Schülerinnen und Schüler auf dieser Stufe besitzen ein gut entwickeltes Verständnis naturwissenschaftlicher Untersuchungen, können ihr Wissen verknüpfen und Situationen kritisch bewerten. Sie können auf Basis ihrer kritischen Analysen auf Beweisen beruhende Erklärungen entwickeln und Aussagen treffen.
IV 559–633 Punkte	Schülerinnen und Schüler auf Stufe IV können mit Situationen und Fragestellungen umgehen, die es erfordern, Rückschlüsse über die Rolle von Naturwissenschaften und Technik zu ziehen. Sie können Erklärungen aus den verschiedenen naturwissenschaftlichen beziehungsweise technologischen Disziplinen auswählen und zueinander in Beziehung setzen, um diese Erklärungen direkt auf alltägliche Lebenssituationen anzuwenden. Schülerinnen und Schüler auf dieser Stufe sind in der Lage, ihre Handlungen zu reflektieren und ihre Entscheidungen auf Basis ihres naturwissenschaftlichen Wissens und der vorhandenen Evidenz zu kommunizieren.
III 485–558 Punkte	Auf Stufe III können Schülerinnen und Schüler eindeutig beschriebene naturwissenschaftliche Fragestellungen in verschiedenen Kontexten erkennen. Sie können Fakten und Wissen auswählen, um naturwissenschaftliche Phänomene zu beschreiben und sind in der Lage, einfache Modelle oder Auswertungsverfahren anzuwenden. Schülerinnen und Schüler auf dieser Stufe können Konzepte aus den verschiedenen naturwissenschaftlichen Disziplinen interpretieren und anwenden. Sie sind in der Lage, Fakten zu verwenden, um kurze Berichte zu verfassen und Entscheidungen zu treffen, die auf naturwissenschaftlichem Wissen beruhen.
II 410–484 Punkte	Schülerinnen und Schüler auf Stufe II besitzen ein angemessenes naturwissenschaftliches Wissen, um in bekannten Kontexten mögliche Erklärungen zu liefern oder aus einfachen Experimenten Schlussfolgerungen zu ziehen. Sie sind in der Lage, in direkten Zusammenhängen zu denken und die Ergebnisse naturwissenschaftlicher Untersuchungen oder technologischer Problemlöseverfahren in ihrer Umgangssprache wiederzugeben.
I 335–409 Punkte	Auf Stufe I ist das naturwissenschaftliche Wissen der Schülerinnen und Schüler derart eingeschränkt, dass es nur auf einige wenige, bekannte Situationen angewendet werden kann. Die Schülerinnen und Schüler auf dieser Stufe sind in der Lage, offensichtliche naturwissenschaftliche Erklärungen zu liefern, die direkt vorliegenden Beweisen zu entnehmen sind.

Fig. 13: Stufen naturwissenschaftlicher Kompetenz in PISA 2012 (Schiepe-Tiska et al. 2013, S. 195)

8 Literaturverzeichnis

COCHRAN, W. G. (1977), *Sampling Techniques, Third edition, John Wiley and Sons*, New York.

DANIEL, J. (2012). *Sampling essentials - practical guidelines for making sampling choices.* Thousand Oaks: Sage.

FAZ a, 2013: *Deutschland im Pisa-Test. Höheres Niveau und ein wenig gerechter.* Zugriff am 21.09.15. Verfügbar unter http://www.faz.net/aktuell/deutschland-im-pisa-test-hoeheres-niveau-und-ein-wenig-gerechter-12693168.html

FAZ b, 2013: *Pisa-Studie. „Deutschland zeigt eine einmalige Entwicklung".* Zugriff am 21.09.15. Verfügbar unter http://www.faz.net/aktuell/politik/pisa-studie-deutschland-zeigt-eine-einmalige-entwicklung-12692486.html

FISCHER, G. H. & MOLENAAR, I. (1995). *Rasch Models: Foundations, recent developments, and applications.* New York: Springer

HEINE, J. H., SÄLZER, C., BORCHERT, L., SIBBERNS, H., & MANG, J. (2013). *Technische Grundlagen des fünften internationalen Vergleichs.* In PRENZEL, M., SÄLZER, C., KLIEME, E., KÖLLER, O. (HRSG.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland (S. 309-346).* Münster: Waxmann

HOHN, K., SCHIEPE-TISKA, A., SÄLZER, C. & ARTELT, C. (2013) *Lesekompetenz in PISA 2012: Veränderungen und Perspektiven in Deutschland.* In PRENZEL, M., SÄLZER, C., KLIEME, E., KÖLLER, O. (HRSG.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland (S. 309-346).* Münster: Waxmann

KALTON, G. (1983). *Introduction to survey sampling.* Newbury Park: Sage.

KISH,L. (1992). *"Weighting for Unequal Pi". Journal of Official Statistics, No.8(2)*, pp.183-200.

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

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2013). *Definitionenkatalog zur Schulstatistik 2012.* Letzter Zugriff am 19.09.2015. Verfügbar unter http://www.kmk.org/statistik/schule/statistische-veroeffentlichungen/definitionenkatalog-zur-schulstatistik.html.

KMK = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2010). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung.* Neuwied: Carl Link.

LEVY, P. S. & LEMESHAW S. (2008). *Sampling of populations - methods and applications* (4. Aufl.). Hoboken, NJ: Wiley & Sons.

LOHR, S.L. (2010), *Sampling: Design and Analysis, Second edition, Pacific, Pacific Grove*, Duxberry.

MASTERS, G. (1982). *A Rasch Model for partial credit scoring. Psychometrika, 47 (2)*, 149?174.

OECD (HRSG.). (2013). *PISA 2012 Ergebnisse, Band I, Was Schülerinnen und Schüler wissen und wie sie dieses Wissen einsetzen können*

OECD (HRSG.). (2014). *PISA 2012, Technical Report*

RASCH, G. W. (1960). *Probabilistic models for some intelligence and attainment tests (Studies in mathematical psychology).* Chicago: The University of Chicago Press.

ROST, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Auflage). Berlin: Hans Huber.

SÄLZER, C. & PRENZEL, M. (2013). *PISA 2012 - Eine Einführung in die aktuelle Studie.* In PRENZEL, M., SÄLZER, C., KLIEME, E., KÖLLER, O. (HRSG.) *PISA 2012 Fortschritte und Herausforderungen in Deutschland* Waxmann Verlag, Münster/New York, München/ Berlin

SÄLZER, C., REISS, K., SCHIEPE-TISKA, A., PRENZEL, M., HEINZE, A. (2013). *Zwischen Grundlagenwissen und Anwendungsbezug: Mathematische Kompetenz im internationalen Vergleich.* In PRENZEL, M., SÄLZER, C., KLIEME, E., KÖLLER, O. (HRSG.) *PISA 2012 Fortschritte und Herausforderungen in Deutschland* Waxmann Verlag, Münster/New York, München/ Berlin

SÄRNDAL, C.-E., B. SWENSSON AND J. WRETMAN (1992), *Model Assisted Survey Sampling, Springer-Verlag*, New York.

SCHIEPE-TISKA, A., SCHÖPS, K., RÖNNEBECK, S., KÖLLER, O. & PRENZEL, M. (2013). *Naturwissenschaftliche Kompetenz in PISA 2012: Ergebnisse und Herausforderungen.* In In PRENZEL, M., SÄLZER, C., KLIEME, E., KÖLLER, O. (HRSG.) *PISA 2012 Fortschritte und Herausforderungen in Deutschland* Waxmann Verlag, Münster/New York, München/ Berlin

SPIEGEL (2014). *Erfindungen: Deutschland ist Patent-Europameister* Zugriff am: 21.09.2015. Verfügbar unter http://www.spiegel.de/wirtschaft/unternehmen/patente-32-000-anmeldungen-kamen-2014-aus-deutschland-a-1020641.html

VAN DER LINDEN, J. W., VELDKAMP, B. P. & CARLSON, J. E. (2004). *Optimizing balanced incomplete block designs for educational assessments.* Applied Psychological Measurement, 28 (5), 317-331.

WUTTKE (2006): *PISA - ein Zufallsgenerator.* Zugriff am: 21.09.15. Verfügbar unter http://www.abafachverband.org/fileadmin/user_upload/user_upload_2007/politik-zeitgeschehen/wuttkepisazufallsgenerator.pdf