universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Zur Stabilität der Itemparameterschätzungen der PISA Studie über die Zeit“

verfasst von / submitted by

Nina Zabransky

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, 2015 / Vienna, 2015

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 298
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Psychologie
Betreut von / Supervisor:	Univ.-Prof. i.R. Dr. Mag. Klaus Kubinger

Danksagung

Ich möchte mich besonders bei Herrn Univ.-Prof. Dr. Mag. Klaus Kubinger für die Unterstützung und Betreuung dieser Diplomarbeit bedanken. Ich danke auch Herrn ao. Univ.-Prof. Dr. Erich Neuwirth für die zur Verfügung gestellten Daten der PISA Studie. Zudem möchte ich mich bei Herrn Mag. Jan Steinfeld für die schnelle Hilfe bei technischen Problemen bedanken.

Mein Dank gilt auch ganz besonders meinen Eltern, meinem Bruder und meinen Freunden für die Unterstützung während des Studiums und die konstruktive Kritik zu dieser Arbeit.

Inhaltsverzeichnis

I. Einleitung

Die PISA Studie (Programme for International Student Assessment) basiert auf dem dichotom logistischen Modell von Rasch (1980). Der Vorteil dieses theoretischen Modells liegt darin, dass es einen internationalen Vergleich von Schulleistungen unter stark variierenden Bedingungen, wie beispielsweise unterschiedliche Bildungssysteme und Sprachen, in verschiedenen Ländern erlaubt. Da in der PISA Studie eine Beschreibung der internationalen Niveauunterschiede von Schulleistungen ermöglicht werden soll, müssen die Items in den verschiedenen Ländern vergleichbar sein. Bei dem verwendeten Verrechnungsmodus der PISA Studie ist die Gültigkeit des Rasch-Modells für die Vergleichbarkeit zwingend notwendig.

Ziel dieser Arbeit ist es daher, die Daten der PISA Studie, die das erste Mal im Jahr 2000 stattfand, auf ihre Rasch-Modell-Konformität zu überprüfen. Zuerst soll auf die PISA Studie näher eingegangen werden, mit Schwerpunkt auf dem zu Grunde liegenden theoretischen Modell und den Methoden der Modellüberprüfung. Weiters werden bereits durchgeführte Studien zur Überprüfung der Rasch-Modell-Konformität bezüglich des Kriteriums *Geschlecht* und *unterschiedliche Länder* sowie über *verschiedene Testzeitpunkte* hinweg beschrieben.

II. Theoretischer Teil

1. Die PISA Studie

Die PISA Studie ist eine von den Mitgliedsstaaten der OECD (englisch für: Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) erstmalig durchgeführte, groß angelegte Untersuchung zur Erfassung der Kompetenzen von SchülerInnen, um den teilnehmenden Staaten vergleichbare Daten zum Stand ihrer Bildungssysteme zur Verfügung zu stellen (OECD, 2012).

> *The OECD Programme for International Student Assessment (PISA) is a collaborative effort among OECD member countries to measure how well 15-year-old students approaching the end of compulsory schooling are prepared to meet the challenges of today's knowledge societies.* (OECD, 2012, S.22)

Dieser Aussage der OECD folgend ist das Ziel der PISA Studie nicht die Überprüfung der Lernerfolge hinsichtlich der Lehrpläne, sondern der Fähigkeit, das Gelernte in lebensnahen Aufgaben anzuwenden. Getestet werden 15 bis 16 jährige SchülerInnen, da sie sich in den meisten OECD Mitgliedsstaaten am Ende ihrer Pflichtschulzeit befinden.

Die PISA Studie fand das erste Mal im Jahr 2000 statt und wurde seither alle drei Jahre wiederholt, zuletzt im Jahr 2012. Bei jeder Untersuchung werden jeweils drei Kompetenzbereiche erhoben: Lesekompetenz (reading literacy), mathematische Kompetenz (mathematical literacy) und naturwissenschaftliche Kompetenz (scientific literacy). Jedes Jahr wird auf einen anderen Bereich der Schwerpunkt gelegt. Im Jahr 2000 stand die Lesekompetenz im Vordergrund, während 2003 die mathematische und 2006 die naturwissenschaftliche Kompetenz genauer untersucht wurde, im Jahr 2009 begann der Zyklus von vorne. Dem Schwerpunkt werden zwei Drittel der Testzeit zugeteilt, um ihn umfangreich zu untersuchen, während die anderen Bereiche nur global erfasst werden. Teilweise werden diese drei Kompetenzbereiche noch erweitert; so wurde beispielsweise im Jahr 2003 zusätzlich die Problemlösefähigkeit erhoben. Des

Weiteren kommen ein Fragebogen für SchülerInnen zum Einsatz, um Informationen zu den Lerngewohnheiten, dem sozialen Hintergrund, der Wohnsituation ect. zu erfassen, sowie ein Schulfragebogen, in dem Hintergrundinformationen zu den Charakteristiken der Schule gesammelt werden (OECD, 2012).
Die PISA Studie wird als Papier-Bleistift Test durchgeführt. Es gibt auch bereits eine auf Computern basierende Testversion, auf diese wird in dieser Forschungsarbeit jedoch nicht näher eingegangen. Die Anzahl der Teilnehmerstaaten steigt kontinuierlich an; während im Jahr 2000 32 Länder begonnen haben (davon 28 OECD Mitgliedsstaaten), waren es bei der Testung 2009 bereits 65 Länder (OECD,2012).

Die Anzahl der Teilnehmerstaaten, sowie das dadurch erforderliche länderübergreifende und somit auch sprachenübergreifende Design, verlangen besondere Maßnahmen, um die Vergleichbarkeit der Daten gewährleisten zu können. Insbesondere muss auf folgende Punkte besonderes Augenmerk gelegt werden:

1. Auf vielfältige operationale Vorgehensweisen, wie verwaltungstechnische Vereinbarungen, Datenerhebung und –verarbeitung sowie durchgehende Qualitätssicherung
2. Auf das Testdesign
3. Auf die Übersetzung in die verschiedenen Landessprachen
4. Auf die Erhebung der Stichprobe, sowohl auf Schulebene als auch auf SchülerInnenebene
5. Auf die Skalierung und die Analyse der Daten mit Hilfe der Item Response Theory (IRT)
6. Auf die Berichterstattung der Daten
   (OECD, 2012)

In den folgenden Kapiteln wird ausschließlich auf die Punkte 2, 3, 4 und 5 eingegangen, da diese für diese Forschungsarbeit relevant sind.

1.1. Das Testdesign

Bei jeder PISA Studie werden mindestens drei Kompetenzbereiche erfasst: Mathematik, Lesen und Naturwissenschaften. Die Items zur Erfassung der Kompetenzen beziehen sich jeweils auf einen bestimmten Stimulus, beispielsweise auf eine Textpassage oder auf ein Diagramm. Diese Items werden in Gruppen, auch Cluster genannt, zusammengefasst, deren Bearbeitungsdauer jeweils 30 Minuten in Anspruch nimmt. Insgesamt stehen pro Testung sieben Cluster mit Items aus dem jeweiligen Schwerpunkt, sowie jeweils drei Cluster für die anderen Kompetenzbereiche zur Verfügung. Da die Bearbeitung aller Items über sechs Stunden in Anspruch nehmen würde, muss jeder Schüler / jede Schülerin nur einen Teil der Aufgaben bearbeiten. Die SchülerInnen bekommen randomisiert eines von 13 möglichen Testheften vorgelegt, in denen sich jeweils vier Cluster befinden und daher insgesamt zwei Stunden an Bearbeitungszeit benötigt wird (OECD, 2014b). Zusätzlich wurde für SchülerInnen mit sonderpädagogischem Förderbedarf eine gekürzte Testversion mit leichteren Items erstellt, deren Bearbeitung nur eine Stunde in Anspruch nimmt (OECD, 2012).

Die Zuordnung der Cluster zu den unterschiedlichen Testheften erfolgt nach einem balancierten unvollständigen Block Design, das im PISA Technical Report der OECD (2014b) folgendermaßen beschrieben wird:

> *Each cluster (and therefore each test item) appears in four of the four-cluster test booklets, once in each of the four possible positions within a booklet, and each pair of clusters appears in one (and only one) booklet. An additional feature of the PISA 2012 test design is that one booklet (booklet 12) is a complete link, being identical to a booklet administered in PISA 2009. (S.30)*

Dieses Design hat zur Folge, dass nicht jeder Schüler / jede Schülerin Aufgaben aus jedem Kompetenzbereich bearbeitet (OECD, 2014b). Die Items werden nicht für jede Untersuchung neu konstruiert, sondern auch aus den vorherigen Testungen wieder verwendet. Zumeist werden ganze intakte Cluster übernommen (OECD, 2014b), teilweise werden aber auch nur einzelne Items aus früheren

Untersuchungen ausgewählt. Diese Items werden *Linking Items* genannt, da sie die Testungen zu unterschiedlichen Zeitpunkten miteinander verbinden (OECD, 2012). Auf diese Linking Items wird zu einem späteren Zeitpunkt noch genauer Bezug genommen (siehe Kapitel 3.4.).

Die Items unterscheiden sich hinsichtlich ihres Formats. Einerseits gibt es klassische Multiple-Choice Aufgaben (MC-Aufgaben), bei denen die SchülerInnen die passende Antwort aus vier oder auch fünf gegebenen Antwortmöglichkeiten wählen müssen, andererseits auch offene Fragen, die entweder eine kurze Antwort erfordern, bestehend aus nur einer Zahl oder einem Wort, oder auch eine ausführlichere Erklärung verlangen. Hilfsmittel wie Taschenrechner dürfen verwendet werden, sollten aber nur erlaubt werden, wenn ihr Gebrauch im normalen Schulalltag selbstverständlich ist (OECD, 2012).

1.2. Die Übersetzung in die verschiedenen Landessprachen

Um die Vergleichbarkeit der Items in verschiedene Sprachen zu gewährleisten, werden besondere Maßnahmen ergriffen. Zuerst werden Items von Testentwicklungsteams vorbereitet; diese basieren einerseits auf vorgeschlagenen Items – im Jahr 2012 wurden von 21 Ländern Items vorgeschlagen – und andererseits auf eigens entwickelten Items. Um gängige Fehler, die in vorangegangenen PISA Studien aufgetreten sind, bei der Übersetzung oder bezüglich kultureller Differenzen zu vermeiden, werden den Testentwicklungsteams einschlägige Informationen und Trainings zur Verfügung gestellt. Die englischen Items werden gleich auf Französisch übersetzt, woraus sich zwei *Ursprungsversionen* ergeben, eine in Englisch und eine in Französisch. Diese Vorgehensweise hat den Vorteil, dass die Übersetzung in die französische Sprache als Probelauf dient, um eventuell auftretende Schwierigkeiten bei der Übersetzung in eine andere Sprache frühzeitig zu entdecken und ihnen so entgegenwirken zu können (OECD, 2012).
Schließlich werden Itemanalysen an Hand der Daten, die bei einem Feldversuch in Ländern mit Französisch bzw. Englisch als Hauptsprache erhoben wurden,

durchgeführt. Werden Unregelmäßigkeiten zwischen den unterschiedlichen Versionen entdeckt, werden die betroffenen Items genau verglichen und gegebenenfalls modifiziert (OECD, 2014b).

Danach findet die nationale Übersetzung in die einzelnen Landessprachen statt. Diese folgt dem Prinzip der „doppelten Übersetzung von beiden Ursprungsversionen". Das bedeutet, es werden jeweils Übersetzungen von zwei unabhängigen Personen vorgenommen, sowohl für die englische als auch für die französische Version, die wiederum von einer dritten Person abgeglichen werden. So entstanden im Jahr 2012 insgesamt 98 nationale Versionen in 46 Sprachen. Die unterschiedlichen nationalen Versionen ergaben sich aus der Tatsache, dass jedes Land die Möglichkeit hat Anpassungen durchzuführen, um den nationalen und kulturellen Besonderheiten ihres Landes gerecht zu werden.
Um die angestrebten hohen Qualitätsstandards der Übersetzungen zu garantieren, wird ein unabhängiges und geschultes Team von Experten eingesetzt, um die nationalen Versionen noch einmal zu kontrollieren und die Übersetzungen gegen die englische und/oder die französische Ursprungsversion abzugleichen (OECD, 2012b).

Einen besonderen Fall stellen die *Linking Items* dar. Das sind jene Items, die bereits in den Jahren zuvor Verwendung fanden und dazu dienen sollen, die Ergebnisse der unterschiedlichen Zeitpunkte miteinander zu verbinden und dadurch vergleichbar zu machen. Um die Übersetzung dieser Linking Items zu verifizieren wird besonders auf die Gleichheit der Formulierung zu jedem Zeitpunkt der Studie Wert gelegt. Ein großes Problem stellt dabei die Tatsache dar, dass in manchen Ländern kleinere Änderungen an den Items vorgenommen wurden, ohne diese genau zu dokumentieren, obwohl bekannt ist, dass sogar kleinste Abweichungen in den Formulierungen zu einem modifizierten Antwortmuster führen können (OECD, 2014b).

1.3. Das Stichprobendesign

Ziel der PISA Studie ist es, eine Aussage über die Fähigkeiten von allen 15 Jahre alten SchülerInnen der teilnehmenden Staaten zu treffen. Da eine Testung aller SchülerInnen nicht möglich ist, muss eine repräsentative Auswahl aus der Zielpopulation getroffen werden. Die Zielpopulation stellt alle Vollzeit- und Teilzeit-SchülerInnen aus einem bestimmten Geburtsjahrgang dar, das bedeutet, dass sowohl 15 wie auch 16 Jahre alte SchülerInnen getestet werden, wenn sie dem entsprechenden Jahrgang zugehören (OECD, 2014b).

In beinahe allen Ländern, außer Russland, wird eine zweistufig geschichtete Zufallsauswahl vorgenommen. Das bedeutet, dass zuerst aus allen Schulen, in denen SchülerInnen der Zielpopulation unterrichtet werden, einige Schulen per Zufall ausgewählt werden. Die Wahrscheinlichkeit, dass eine Schule ausgewählt wird, ist umso größer, je größer der Anteil an eingeschriebenen SchülerInnen aus der Zielpopulation ist. Im zweiten Schritt werden aus den selektierten Schulen per Zufall SchülerInnen ausgewählt, die an der PISA Studie teilnehmen. Aus jedem teilnehmenden Land müssen mindestens 150 Schulen ausgewählt werden, aus denen jeweils 35 SchülerInnen per Zufall selektiert werden; so ergibt sich mindestens eine Stichprobe von 4500 SchülerInnen pro Land. Ausnahmen stellen kleinere Länder, wie Liechtenstein, dar (OECD, 2014b).

Vor der Auswahl der StudienteilnehmerInnen werden die Schulen an Hand bestimmter Variablen zu Gruppen zusammengefasst, auch Schichten genannt, die in weiterer Folge als unabhängig angesehen werden. Die Definition der Schichten hängt von den Ländern ab. In Österreich werden vorrangig die unterschiedlichen Schultypen als Schichtungskriterium herangezogen, in anderen Ländern kommen auch andere Kriterien zum Tragen, wie beispielsweise der Grad der Verstädterung oder der Anteil von Minderheiten (OECD, 2014b). Grossmann, Ledl, Neuwirth und Steiner (2006, S.13) geben zu bedenken: „Als Folge davon ist aus statistischer Sicht das Studiendesign in den einzelnen Ländern bereits nicht mehr ganz identisch."

Damit jeder zufällig ausgewählte Schüler oder jede Schülerin die entsprechende Anzahl an SchülerInnen aus der Gesamtpopulation hinsichtlich der interessierenden Variablen auch richtig repräsentiert, wird bei den Berechnungen der Ergebnisse ein komplexes Gewichtungsschema angewendet (OECD, 2014b), auf das hier jedoch nicht näher eingeganen werden soll.

2. Das Rasch-Modell

Das Rasch-Modell ist ein im statistischen Sinn stichprobenunabhängiges Modell (z. B. Kubinger, 2009), das zum ersten Mal im Jahre 1960 von Georg Rasch (Rasch, 1980) publiziert wurde. Es beschreibt die Wahrscheinlichkeit, dass eine Person v mit der Fähigkeit ξv, das Item *i*, mit der Schwierigkeit σi, löst (+), folgendermaßen:

$$P\left(+|\xi_v, \sigma_i\right) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}$$

Abbildung 1. Formel des Rasch-Modells (z.B. Kubinger, 2009, S.89)

Das bedeutet, dass die Wahrscheinlichkeit, ob eine Person ein Item richtig beantworten kann, neben dem Zufall, nur von der Personenfähigkeit und der jeweiligen Itemschwierigkeit abhängt.

2.1. Eigenschaften des Rasch-Modells

Das Rasch-Modell weist besondere Eigenschaften/Voraussetzungen auf, die erfüllt sein müssen, damit die Gültigkeit angenommen werden kann. Diese Eigenschaften sollen im Folgenden kurz beschrieben werden.

2.1.1. Suffizienz

Suffiziente Statistik bedeutet, dass alle notwendigen Informationen zur Berechnung der Personenfähigkeit bereits in der Anzahl der gelösten Aufgaben enthalten sind. Damit erhalten die Personen mit demselben Score, also derselben Anzahl an gelösten Items, auch dieselbe Personenfähigkeit und zwar unabhängig

davon, welche Items sie im Speziellen gelöst haben (Fischer, 1974). In vielen psychologischen Tests wird „die Anzahl der gelösten Items“ als Grundlage zur Schätzung der Personenfähigkeit herangezogen. Dabei wird jedes Item dichotom verrechnet, also entweder als „gelöst“ oder „nicht gelöst“. Damit diese Art der Verrechnung zu einem fairen Vergleich der Personenfähigkeiten führt, muss zwingend das Rasch-Modell gelten (Kubinger, 2005).

2.1.2. Eindimensionalität

Eindimensionalität bedeutet, dass alle Items eines Tests dieselbe Eigenschaft messen. Die Lösungswahrscheinlichkeit aller Items hängt von nur einer Fähigkeitsdimension der Person ab (Fischer, 1974). Beispielsweise soll ein Test, der die mathematische Kompetenz messen soll, auch nur diese erfassen und nicht etwa zusätzlich Sprachfähigkeiten.

2.1.3. Lokale stochastische Unabhängigkeit

Lokale stochastische Unabhängigkeit bedeutet, dass die Wahrscheinlichkeit ein Item zu lösen, nur von den Eigenschaften dieses Items und der Fähigkeit der Person abhängt und nicht davon, welche anderen Items bereits gelöst wurden oder erst noch gelöst werden. Das bedeutet, dass es während der Bearbeitung der Items zu keinen Lernprozessen kommen darf und die Lösung eines Items nicht auf dem Resultat eines anderen Items aufbauen soll (Fischer, 1974).

2.1.4. Spezifische Objektivität / Stichprobenunabhängigkeit

Stichprobenunabhängigkeit bedeutet, dass sich das Ergebnis der Parameterschätzungen nicht systematisch ändert, wenn man von einer Stichprobe zu einer anderen Stichprobe aus der definierten Population übergeht. Umgekehrt ist es ebenso irrelevant, welche Items aus dem vorhandenen Itempool

vorgegeben werden, um die Personenfähigkeit zu bestimmen (Fischer, 1974). Wenn sich die Itemparameterschätzungen von wenigstens einem Item zwischen unterschiedlichen Subgruppen von Personen aus der untersuchten Population (wie etwa Männer und Frauen, Personen aus unterschiedlichen Ländern) unterscheiden, dann spricht man von einem *Differential Item Functioning (DIF).* Weist ein Test DIF auf, gilt das Rasch-Modell nicht, da gleiche Testwerte nicht gleiche Leistungen wiederspiegeln (Kubinger, 2015).

2.1.5. Streng monoton steigende Itemcharachteristikkurven

Die Itemcharakteristikkurve (ICC) stellt die Wahrscheinlichkeit, ein Item zu lösen, in Abhängigkeit von der Personenfähigkeit graphisch dar. Wenn das Rasch-Modell gilt, zmüssen alle ICCs streng monoton ansteigen. Das bedeutet, dass die Wahrscheinlichkeit, ein Item zu lösen mit zunehmender Personenfähigkeit durchgehend ansteigt und niemals konstant bleibt oder wieder absinkt. Zudem verlaufen die ICCs aller Items parallel und können sich niemals überschneiden. Ist dies nicht der Fall, dann ist eine Verletzung des Rasch-Modells anzunehmen (Fischer, 1974).

2.2. Überprüfung der Modellgültigkeit

Um die zu Grunde liegenden Annahmen des Rasch-Modells empirisch zu überprüfen, werden bestimmte Verfahren bzw. Tests angewendet. Treffen die Annahmen für einen vorliegenden Datensatz nicht zu, kann davon ausgegangen werden, dass eine Modellverletzung vorliegt (Fischer, 1974). Im Folgenden werden die für diese Forschungsarbeit relevanten Modellkontrollen kurz dargestellt.

2.2.1. Die graphische Modellkontrolle

Die graphische Modellkontrolle basiert auf der Vorraussetzung der Stichprobenunabhängigkeit der Parameterschätzungen des Rasch-Modells. Um diese zu überprüfen wird die Stichprobe nach einem geeigneten Kriterium (etwa Anzahl gelöster Aufgaben, Alter) unterteilt. So kann überprüft werden, ob die für die Teilstichproben getrennt berechneten Parameter überzufällig voneinander abweichen. Dies ist in einer graphischen Abbildung leicht erkennbar; gilt das Rasch-Modell, stimmen die Werte überein und liegen in einem Koordinatensystem auf einer 45°-Geraden durch den Ursprung (Fischer, 1974).

2.2.2. Der Andersen-Likelihood-Ratio-Test

Der Likelihood-Ratio-Test (LRT; auch Likelihood-Quotienten-Test) von Andersen (1973) basiert ebenfalls auf dem Vergleich verschiedener Teilgruppen. Hierfür werden Maximum-Likelihood-Schätzungen für die Itemparameter sowohl für die Gesamtgruppe, als auch für jede Teilgruppe berechnet. Wenn das Rasch-Modell gilt, dürfen sich die Werte aus der Gesamtgruppe und die der Teilgruppen nicht signifikant voneinander unterscheiden. Die Stichprobe kann an Hand unterschiedlicher Teilungskriterien in Gruppen aufgegliedert werden. Es kann ein internes Teilungskriterium verwendet werden, wie etwa die Anzahl der gelösten Aufgaben oder es wird ein Außenkriterium herangezogen, wie beispielsweise das Geschlecht (Fischer, 1974).

3. Das Rasch-Modell in der PISA Studie

Das Rasch-Modell bietet sich als Grundmodell für die PISA Studie an, da diese den Anspruch stellt, die Leistungen von SchülerInnen aus unterschiedlichsten Ländern und Sprachen miteinander zu vergleichen und dies auf einem möglichst Ressourcen schonenden Weg durchzuführen.

In weiterer Folge wird das Skalierungsmodell der PISA Studie sowie dessen Überprüfung genauer vorgestellt. Zuerst wird erklärt, wie die Personenfähigkeitsparameter geschätzt werden (siehe Kapitel 3.1.), anschließend wird beschrieben, wie die Itemparameter geschätzt und überprüft werden (siehe Kapitel 3.2.).

3.1. Fähigkeitsparameter und Skalierung der PISA Studie

Der PISA Studie liegt die Idee einer eindimensionalen, normalverteilten, latenten Fähigkeit zu Grunde, auf welche durch die Testwerte geschlossen wird. Die Auswertung der dichotomen Items (richtig/falsch) erfolgt mittels Rasch-Modell. Die Schätzung der Fähigkeitsparameter der einzelnen SchülerInnen wird mittels der Maximum-Likelihood-Methode durchgeführt (Grossmann et al., 2006). Damit die Testwerte die latente Fähigkeit richtig abbilden ist ein Skalierungsmodell notwendig. Das Skalierungsmodell beschreibt daher nach Kubinger (2009) die Verrechnungsvorschriften, um die aus den Testwerten berechneten Verhaltensrelationen adäquat abzubilden.

Die Grundlage dieser Verrechnung ist ein Regressionsmodell (Grossmann et al., 2006). Dabei wird davon ausgegangen, dass die Fähigkeiten der SchülerInnen von mehreren Variablen abhängen (OECD, 2014b). Laut Kreiner und Christensen (2014) sind das folgende:

> *The PISA scaling model contains items (Y1, . . . , Y28), a unidimensional latent variable Θ, the variable country C, and other exogenous variables X including gender, age, and grade of student, education and occupation of parents, type and size of schools, class sizes, and many other variables.* (S.214)

Mit Hilfe dieses Regressionsmodells wird also die Verteilung der latenten Variable – Lesekompetenz, mathematische und naturwissenschaftliche Kompetenz – innerhalb der zu untersuchenden Population in Abhängigkeit der oben angeführten *conditioning variables,* wie Geschlecht, Schule oder Ausbildungsniveau der Eltern, beschrieben (Kreiner & Christensen, 2014).
Die Verteilung der latenten Variable aller SchülerInnen, die sich in den conditioning variables gleichen, wird *Priorverteilung* genannt. Die conditioning variables führen zu einer systematischen Veränderung im Antwortmuster, welche die Vergleichbarkeit der Ergebnisse beeinträchtigt. Die *Posteriorverteilung* gibt die Verteilung der latenten Variable unter Berücksichtigung der bekannten Testantwortmuster und conditioning variables an (Grossmann et al., 2006).

Plausible Values

Aus den Posteriorverteilungen der SchülerInnenfähigkeiten werden die länderspezifischen Durchschnittsleistungen berechnet. Da aber auf Grund des bereits vorgestellten Testdesigns (siehe Kapitel 1.1.) nicht jeder Schüler / jede Schülerin Items aus allen drei Kompetenzbereichen bearbeitet, können deren Fähigkeiten nicht für jeden Kompetenzbereich berechnet werden. Fehlende Werte können jedoch aus den bearbeiteten Items geschätzt werden. Diese Ableitung wird bei PISA mit Hilfe der *Plausible Values* (PVs) durchgeführt (OECD, 2014b). Hierzu werden für jeden Schüler / jede Schülerin aus der Posteriorverteilung für jeden Kompetenzbereich fünf plausible Werte (PVs) für die Personenfähigkeit generiert. Es handelt sich also um statistisch generierte Zufallszahlen, die angeben welche Fähigkeitsparameter bei gegebenen Hintergrundvariablen und Testantworten wahrscheinlich sind (Grossmann et al., 2006). Es ist jedoch wichtig zu bedenken, dass PVs keine wahren Testwerte darstellen, sondern nur statistisch generierte Zufallszahlen (OECD, 2014b). Das bedeutet, sie sind nicht geeignet, um Ergebnisse von Einzelpersonen darzustellen,

erlauben aber eine hinreichend genaue Schätzung von Gruppenmittelwerten und Gruppenvarianzen, also die Berechnung der länderspezifischen Durchschnittsleistungen der SchülerInnen (Deutsches PISA-Konsortium, 2003).

3.2. Itemparameter und Modellüberprüfung

Zur korrekten Interpretation der Leistungen der SchülerInnen muss angenommen werden, dass die Voraussetzungen für das Rasch-Modell gegeben sind, beispielsweise, dass die Items lokal stochastisch unabhängig sind und kein DIF aufweisen (Kreiner & Christensen, 2014). Sollte eine Voraussetzung des Modells nicht erfüllt sein, dann wäre der Verrechnungsmodus, nämlich die Anzahl gelöster Items als Testergebnis heranzuziehen und somit die Ergebnisse der PISA Studie, nicht fair vergleichbar und interpretierbar. Um zu überprüfen, ob die Items der PISA Studie auch wirklich die Voraussetzungen des Rasch-Modells erfüllen, werden verschiedene Werte berechnet und analysiert. Neben der deskriptiven Beschreibung der Items, wird besonderes Augenmerk auf die folgenden drei Punkte gelegt: Berechnung der *Infit-Teststatistik*, der *Diskriminationskoeffizienten* und der *item-by-country Interaktion* (siehe Kapitel 3.2.1. - 3.2.3.) (OECD, 2014b). Die Berechnungen dieser Werte finden sowohl auf nationaler als auch auf internationaler Ebene statt und werden anschließend miteinander verglichen. Große Unterschiede zwischen den nationalen und den internationalen Werten zeigen, dass sich ein Item in einem Land im Vergleich zu den übrigen Ländern anders verhält (Schwantner & Schreiner, 2010).

Die Itemparameter werden für jedes Land einzeln mittels der marginalen Maximum-Likelihood-Schätzung (MML) berechnet (Kreiner & Christensen, 2014). Es wird angenommen, dass die teilnehmenden SchülerInnen einer normalverteilten Stichprobe entstammen. Für die MML-Schätzungen werden die ungewichteten Daten verwendet und der Mittelwert der Itemparameter wird in jedem Land auf null festgelegt (OECD, 2012). Auch für die internationale Item-Kalibrierung werden MML-Schätzungen für die Itemparameter durchgeführt. Hierfür wird nicht die gesamte Stichprobe aller Länder herangezogen, sondern

nur eine Teilstichprobe. Diese setzte sich im Jahr 2012 aus allen Ländern zusammen, die bei der Papierform von PISA teilgenommen haben. Genauer wurden 500 SchülerInnen zufällig von jedem der 63 teilnehmenden Länder ausgewählt, insgesamt ergab sich so das *international calibration sample* mit insgesamt 31 500 SchülerInnen (OECD, 2014b).

3.2.1. Infit-Teststatistik

Da das Rasch-Modell für jeden Schüler / jede Schülerin die Wahrscheinlichkeit für das Erreichen der verschiedenen Scores beschreibt, können die durch das Modell vorhergesagten Werte mit den beobachteten Werten verglichen werden. Dabei wird die Summe der quadrierten unstandardisierten Residuen durch die Summe der Residuenvarianzen dividiert (Kreiner & Christensen, 2014, zitiert nach Smith, 2004). Laut der OECD (2014b) ist es dadurch möglich, die erwarteten Werte mit den tatsächlich beobachteten Werten eines Items über alle SchülerInnen hinweg zu vergleichen. Werte um 1 werden angestrebt, Werte innerhalb des Intervalls [0,8 ; 1,2] sprechen für einen ausreichenden Fit der Daten mit dem Modell. Liegen die Werte über 1,2 weist das darauf hin, dass die Daten mehr Variabilität aufweisen als erwartet, liegen sie unter 0,8, weisen sie weniger auf als erwartet (OECD, 2014b).

3.2.2. Diskriminationskoeffizient

Für jedes Item wird ein Diskriminationskoeffizient berechnet. Hierzu wird eine Produkt-Moment-Korrelation zwischen dem SchülerInnen-Score und dem aufsummierten Score jener Items, die dem zu überprüfenden Item im Kompetenzbereich (z.B. Lesen) gleichen, gebildet (OECD, 2014b). Er beschreibt, in wieweit die Modellanpassung zwischen SchülerInnen mit hohen Testwerten und SchülerInnen mit niedrigen Testwerten unterscheidet, also wie hoch ihre Trennschärfe ist. Der Diskriminationskoeffizient soll Werte zwischen 0,3 und 0,7 aufweisen, Items mit einem Koeffizienten unter 0,2 zeigen eine geringe

Trennschärfe an, negative Werte lassen auf deutliche Probleme bei dem Item schließen (Schwantner & Schreiner, 2010).

Für Multiple-Choice Aufgaben und solche, die nur eine kurze Antwort verlangen, wird der Punkt-Biseriale Diskriminationsindex gebildet. Er wird für jede Antwortkategorie eines Items berechnet, indem die aufsummierten Scores jener SchülerInnen, die diese Kategorie gewählt haben, mit jenen, die diese Kategorie nicht gewählt haben, verglichen werden (OECD, 2014b). Wenn die untersuchte Antwortkategorie die richtige Antwort darstellt, sollte der Diskriminationskoeffizient mindestens einen Wert über 0,20 erreichen (Ebel & Frisbie, 1986).

3.2.3. Item-by-country Interaktion

Für jedes Land werden national spezifische Itemparameterschätzungen durchgeführt, die in weiterer Folge in die internationale Datenbank aufgenommen werden. Von besonderem Interesse ist hierbei die Konsistenz der Schätzungen über die Länder hinweg. Alle Items sollen in allen Ländern dieselbe Schwierigkeit (zuzüglich des Standardfehlers) aufweisen und somit innerhalb des Konfidenzintervalls liegen. Trifft dies nicht zu, wird nicht in allen Ländern dieselbe latente Fähigkeit gemessen (OECD, 2014b).

Die berechneten Werte werden in nationalen Reporten an die einzelnen Teilnehmerstaaten übermittelt, die in weiterer Folge Urteile darüber fällen, wie mit den Items in den jeweiligen Ländern weiter verfahren werden soll. Für diese Entscheidungen sind besonders die angestrebten Grenzwerte der Infit-Teststatistik, der Diskriminationskoeffizienten und der item-by-country Interaktion wichtig. Wenn die psychometrischen Charakteristiken eines Items in mehr als zehn Ländern nicht zufriedenstellend sind, ist es möglich das Item komplett von den weiteren Analysen auszuschließen. Diese Items werden *dodgy items* genannt. Sind die psychometrischen Charakteristiken nur in einem Land unzureichend, aber in allen anderen ausreichend, dann kann dieses Item auch

nur in diesem speziellen Land von der Datenanalyse ausgenommen sein (OECD, 2012).

3.3. Zusammenfassung: Berechnung der Länderscores

Zuerst findet die Berechnung der Itemparameter auf nationaler Ebene für jedes Land einzeln statt, danach werden die Itemparameter für das international calibration sample, also eine Teilstichprobe aller Länder, geschätzt. Zur Überprüfung der Gültigkeit des Rasch-Modells werden für jedes Item die Infit-Teststatistik, die Diskriminationskoeffizienten und die item-by-country Interaktion berechnet. Auf Basis der Ergebnisse dieser Berechnungen wird entschieden welche Items den Kriterien des Rasch-Modells entsprechen und welche für die weiteren Berechnungen ausgeschlossen werden müssen. In weiterer Folge wird mittels Regressionsmodell die Posteriorverteilung der Fähigkeiten innerhalb der Population in Abhängigkeit der conditioning Variables und des Testergebnisses beschrieben. Aus den Posteriorverteilungen werden für jeden Schüler / jede Schülerin für jeden Kompetenzbereich fünf Plausible Values (PVs) berechnet. Aus diesen PVs werden in weiterer Folge die mittleren Fähigkeiten der SchülerInnen eines Landes berechnet (OECD, 2012b).

3.4. Linking PISA

Um die so gewonnenen Ergebnisse der PISA Studie über mehrere Jahre hinweg vergleichen zu können, werden Linking Items eingesetzt. Das sind Items, die bereits in den Jahren zuvor vorgegeben wurden und somit einen direkten Vergleich der Itemparameter erlauben. Beispielsweise wurden im Jahr 2000 im Kompetenzbereich Lesen insgesamt 129 Items vorgegeben, wovon dieselben 26 Items auch in den Jahren 2003, 2006 und 2009 wieder verwendet wurden (OECD, 2012).

Um mögliche Unterschiede in der Schwierigkeit der Items zu den unterschiedlichen Zeitpunkten auszugleichen, werden folgende Schritte unternommen:
Zuerst werden die Itemparameter des international calibration sample geschätzt, beispielsweise von dem Jahr 2012. Diese werden mit den Schätzungen der letzten PISA Studie aus dem Jahr 2009 verglichen, um eine *shift Konstante* zu berechnen, die die Itemparameterschätzungen von 2012 und 2009 auf denselben Mittelwert setzt. Die Personenfähigkeiten werden mit den Daten aus der Studie 2012 geschätzt und danach mit der shift Konstante transformiert, um die Jahrgänge vergleichbar zu machen.
Da die Berechnung der shift Konstante von den ausgewählten Linking Items abhängt, würde sie andere Werte annehmen, wenn andere Linking Items ausgewählt worden wären. Diese Unsicherheit wird durch den *linking error* ausgedrückt. Er muss berücksichtigt werden, wenn Daten aus unterschiedlichen Jahren miteinander verglichen werden sollen (OECD, 2014b).

3.5. Bisherige Ergebnisse aus der Literatur

Es wurden bereits mehrere Studien durchgeführt, die zum Ziel hatten, die Qualität und Voraussetzungen des methodischen Modells der PISA Studie zu überprüfen. Hier sollen Studien mit Bezügen zu dieser Arbeit angeführt werden.

Kreiner und Christensen (2014) haben überprüft, ob die Items im Kompetenzbereich Lesen der PISA Studie 2006 dem Rasch-Modell entsprechen. Dabei wurde nur jenes Testheft herangezogen, das alle Items des Kompetenzbereichs Lesen enthielt. Die Autoren führten einerseits dieselben Analyseschritte durch, die bei PISA unternommen wurden, um die Modellgültigkeit zu überprüfen, andererseits ergänzten sie diese noch um weitere Schritte. So haben sich die Autoren dazu entschieden, auch den Andersen-Likelihood-Ratio-Test (Andersen, 1973) anzuwenden, da in der PISA Studie kein globaler Test zur Überprüfung der Modell-Passung durchgeführt wird. Ihre Berechnungen basieren daher auf den bedingten Maximum-Likelihood-

Parameterschätzungen (CML), im Gegensatz zu den in der PISA Studie verwendeten MML-Schätzungen. Es wurde ein internes Teilungskriterium verwendet (Anzahl gelöster Aufgaben) und zwar sowohl für den gesamten Datensatz als auch für jedes einzelne Land. Sie kamen zu dem Schluss, dass die Rasch-Modellgültigkeit nicht gegeben ist, weder für den gesamten Datensatz noch in den einzelnen Ländern. Es waren auch bei beinahe allen Items die Itemfit Statistiken unzureichend, ebenso zeigten sich starke Hinweise auf DIF.
Man muss hier allerdings beachten, dass es sich – trotz der Einschränkung auf Items von nur einem Testheft – um eine sehr große Stichprobe handelt und daher eine Ablehnung der Modellgültigkeit zu erwarten ist (Adams, 2011). Kreiner und Christensen (2014) kommen dennoch zu dem Schluss, dass die Beweise, die gegen die Gültigkeit des Rasch-Modells sprechen, überwältigend sind.

Wetzel und Carstensen (2013) analysierten ebenfalls Items der PISA Studie aus dem Kompetenzbereich Lesen für die Jahre 2000 und 2009. Ihr Ziel war es, die Daten hinsichtlich Messbeständigkeit zu untersuchen, da diese eine wichtige Vorraussetzung darstellt, um Trends aus den Daten der PISA Studie abzuleiten. Ihre Stichprobe bestand aus SchülerInnen von 59 Schulen in Deutschland. Diese bearbeiteten im Jahr 2009 zum einen Teil Testhefte der PISA Studie 2009 und zum anderen Teil Testhefte der PISA Studie 2000. Diese 59 Schulen hatten auch schon im Jahr 2000 an der PISA Studie teilgenommen, somit waren Daten sowohl von zwei unterschiedlichen Zeitpunkten, als auch von zwei unterschiedlichen PISA Instrumenten vorhanden. Die Berechnungen wurden, wie in der PISA Studie, mittels ConQuest (Wu, Adams, Wilson & Haldane, 2007) durchgeführt. Die Analysen ergaben, dass sich einige der Linking Items der PISA Studie zwischen den Zeitpunkten 2000 und 2009 im Kompetenzbereich Lesen in ihrer Schwierigkeit deutlich unterscheiden und somit die Messbeständigkeit nicht gegeben ist.

Allerup (2007) untersuchte, ob sich die Itemschwierigkeiten der PISA Studie aus den Jahren 2000 und 2003 in Bezug auf das Geschlecht, die ethnische Gruppenzugehörigkeit und die Zeitpunkte unterscheiden. Die untersuchten Daten bestanden einerseits aus einer repräsentativen Stichprobe der PISA Daten aus den Jahren 2000 und 2003, andererseits aus Daten von allen öffentlichen Schulen in Kopenhagen sowie einer zusätzlichen Stichprobe aus Schulen mit

hoher Zahl an SchülerInnen aus unterschiedlichen ethnischen Gruppen. Um die Unterschiede in den Itemschwierigkeiten im Kompetenzbereich Lesen in Bezug auf die zwei Zeitpunkte (Jahre 2000 und 2003) beziehungsweise auf das Geschlecht (Jahr 2003) zu untersuchen, wurden 22 Linking Items analysiert. Zur Untersuchung der unterschiedlichen ethnischen Gruppen wurden ausschließlich Daten aus Dänemark (Jahr 2000) ausgewertet. Der Autor kommt zu dem Schluss, dass die Itemschwierigkeiten der PISA Studie sowohl in Bezug auf die zwei Zeitpunkte, als auch auf das Geschlecht und die ethnischen Gruppen, nicht gleich sind.

III. Empirischer Teil

4. Methode

4.1. Ziel dieser Forschungsarbeit

Da in der PISA Studie eine Beschreibung der internationalen Niveauunterschiede von Schulleistungen ermöglicht werden soll, müssen die Items in den verschiedenen Ländern vergleichbar sein. Bei dem verwendeten Verrechnungsmodus der PISA Studie ist die Gültigkeit des Rasch-Modells für die Vergleichbarkeit zwingend notwendig. Es konnte bereits gezeigt werden, dass Voraussetzungen des Rasch-Modells verletzt wurden. Kreiner und Christensen (2014) haben für den Kompetenzbereich Lesen im Jahr 2006 nachgewiesen, dass beinahe alle Daten gegen die Gültigkeit des Rasch-Modells sprechen. Zudem kam Allerup (2007) in seiner Studie zu dem Schluss, dass die Itemschwierigkeiten der PISA Studie im Jahr 2000 Differenzen in Bezug auf das Geschlecht aufweisen.

Allgemein ist bei der Modellüberprüfung mittels großer Stichproben zu kritisieren, dass bereits kleine Unterschiede statistisch signifikant werden (Kubinger, 2005). Es stellt sich die Frage, ob die berechneten signifikanten Modellabweichungen (z.B. Kreiner und Christensen, 2014) auch praktisch relevant sind. Die vorliegende Arbeit setzt sich zum Ziel, die Gültigkeit des Rasch-Modells an Hand der Daten der PISA Studie der letzten Jahre (2000 bis 2012) nicht auf statistisch signifikante, sondern auf praktisch relevante Abweichungen zu überprüfen.

Forschungsfragen:
Zeigen die in der PISA Studie verwendeten Items eine praktisch relevante Abweichung vom Rasch-Modell:

1. über die Zeitpunkte hinweg?
2. bezüglich des Geschlechts?
3. über die verschiedenen beteiligten Länder hinweg?

Die genaue Vorgehensweise wird in den folgenden Kapiteln erörtert.

4.2. Durchführungsplan

Die Auswertung der PISA Daten soll mit dem Zusatzprogramm eRm (*extended Rasch modeling*; Mair, Hatzinger, & Maier, 2014, Version 0.15-5) der Open Source Software R (R Development Core Team, 2015, Version 3.2.2) erfolgen, das eigens für Berechnungen mit der Item Response Theory entwickelt wurde.

4.2.1. Stichprobe

Um die Konformität des Rasch-Modells zu überprüfen, soll eine Stichprobe bestehend aus 1 905 147 SchülerInnen ausgewertet werden. Die Daten wurden im Zuge der PISA Studie der letzten Jahre (2000, 2003, 2006, 2009, 2012) erhoben und von Herrn ao. Univ.-Prof. i.R. Dr. Erich Neuwirth zusammengetragen und freundlicherweise zur Verfügung gestellt.

4.2.2. Praktische Relevanz

Das Signifikanzniveau für den Andersen-Likelihood-Ratio-Test wird auf $\alpha = 0{,}01$ festgesetzt. Da es wahrscheinlich ist, bereits auf Grund der ungewöhnlich großen Stichprobe, signifikante Ergebnisse zu erhalten (Kubinger, 2005), werden für diese Arbeit nicht signifikante, sondern praktisch relevante Abweichungen als Kriterium zur Überprüfung des Rasch-Modells herangezogen. Somit wird auf eine Bonferroni-Korrektur des Signifikanzniveaus verzichtet.
Um einen praktisch relevanten Unterschied bei den Itemparameterschätzungen zwischen den unterschiedlichen Teilgruppen festzustellen, wird – wie bei Kubinger (2005) vorgeschlagen – eine graphische Modellkontrolle durchgeführt; eine Abweichung von der 45°-Geraden kann als praktisch relevant angesehen werden, wenn die Differenz der Itemparameterschätzungen aus zwei Teilstichproben mehr als ein Zehntel der Spannweite der Parameterschätzungen beträgt (Goethals, 1994). Weichen Items um mehr als ein Zehntel der Spannweite

von der 45°-Geraden ab, kann davon ausgegangen werden, dass diese Items nicht Rasch-Modell konform sind.

4.2.3. Vorgehensweise

Da beim Rasch-Modell eine Abweichung bei einem Subset von Items gleichzeitig eine Modellungültigkeit des kompletten Itemsets impliziert (Kreiner & Christensen, 2014), sollen nur jene Items für die Berechnungen ausgewählt werden, die zu jedem Testzeitpunkt vorgegeben wurden (Linking Items). Des Weiteren sollen nur Daten jener Staaten ausgewählt werden, die zu allen Zeitpunkten an der PISA Studie teilgenommen haben. Es wird ausschließlich der Kompetenzbereich Lesen (reading literacy) untersucht, da dieser im Jahr 2000 den ersten Schwerpunkt der PISA Studie darstellte und somit die meisten Linking Items ab dem Jahr 2000 aufweist (OECD, 2012).

Die Auswertung der ausgewählten Daten soll in mehreren Schritten erfolgen:
Zuerst sollen die Itemparameter mittels der bedingten Maximum-Likelihood-Methode (CML) für jeden Erhebungszeitpunkt einzeln (2000, 2003 ect.) geschätzt werden (Kreiner & Christensen, 2014). Danach soll die Modellgültigkeit für jeden Erhebungszeitpunkt einzeln (2000, 2003 ect.) mittels Andersen-Likelihood-Ratio-Test (LRT) überprüft werden. Darüber hinaus sollen die Itemparameter für jedes Land geschätzt werden, um auch diese mittels LRT zu überprüfen, sowohl mit dem Teilungskriterium *Geschlecht*, als auch mit dem internen Teilungskriterium *Anzahl gelöster Aufgaben* (Rohwert kleiner gleich Median vs. Rohwert größer Median).

Im nächsten Schritt sollen jene Länder, die im LRT statistische Signifikanz aufweisen, genauer betrachtet werden. Hierfür soll die graphische Modellkontrolle durchgeführt werden, um festzustellen, ob Items in einem praktisch relevanten Maß vom Rasch-Modell abweichen (siehe Kapitel 4.1.2.). Das bedeutet, wenn keine Items eine relevante Abweichung aufweisen, kann davon ausgegangen werden, dass – trotz statistisch signifikantem Ergebnis im LRT – nach praktisch relevanten Kriterien die Gültigkeit des Rasch-Modells

angenommen werden kann. Weisen Items eine relevante Abweichung auf, kann im Gegenzug davon ausgegangen werden, dass auch nach praktisch relevanten Kriterien die Gültigkeit des Rasch-Modells nicht angenommen werden kann.

Für den Fall, dass zu allen Erhebungszeitpunkten (2000, 2003 ect.) oder in allen Ländern nach praktisch relevanten Kriterien von der Gültigkeit des Rasch-Modells ausgegangen werden kann – bezogen auf die Teilungskriterien Anzahl gelöster Aufgaben und Geschlecht – soll noch ein weiterer Schritt durchgeführt werden. Die Rasch-Modell-Gültigkeit soll über alle Zeitpunkte gemeinsam (2000-2012) bzw. über alle Länder gemeinsam im Sinne einer Äquivalenzprüfung erneut überprüft werden. Falls die Gültigkeit des Rasch-Modells nach praktisch relevanten Kriterien nicht zu allen Erhebungszeitpunkten oder in allen Ländern angenommen werden kann, soll auf diesen Schritt verzichtet werden, da – wie bereits erwähnt – eine Abweichung vom Rasch-Modell bei einem Teil der Daten bereits eine Modellungültigkeit bei dem kompletten Datensatz impliziert (Kreiner & Christensen, 2014).

4.3. Durchführung der Auswertung

Die Durchführung der Auswertung konnte großteils wie geplant stattfinden. Um die sehr große Stichprobe einzugrenzen, wurden jene Länder ausgewählt, die zu allen Durchführungszeitpunkten an der PISA Studie teilgenommen hatten. Die weiteren Schritte wurden, wie geplant, auf den Kompetenzbereich Lesen beschränkt. Es wurden jene Items ausgewählt, die zu allen Zeitpunkten vorgegeben worden sind; hierbei konnten jedoch nur Items aus den Jahren 2000 bis 2009 ausgewählt werden, da im Jahr 2012 nur noch 3 Items mit den vorherigen Testzeitpunkten übereinstimmten. Insgesamt ergab sich so ein Datenfile bestehend aus 35 Ländern, 26 Items und 1 345 279 SchülerInnen.

Da die Stichprobe immer noch sehr groß war, wurde die Analyse, wie bei Kreiner und Christensen (2014), auf vollständige Datensätze begrenzt. Personen mit fehlenden (*missing, not reached*) oder ungültigen (*invalid*) Antworten wurden ebenso von der Datenanalyse ausgeschlossen wie Personen mit fehlenden

Angaben zu ihrem Geschlecht (*unknown*). Um die Itemparameter schätzen zu können, mussten auch noch jene Personen ausgeschlossen werden, die nur eines der Items beantwortet hatten. Des Weiteren wurde die Analyse auf dichotome Items begrenzt, wodurch sich schließlich der zu analysierende Datensatz mit 34 Ländern, 20 Items und 315 072 SchülerInnen ergab.

4.4. Deskriptive Beschreibung der Stichprobe

4.4.1. Ursprungsstichprobe

Die ursprüngliche Stichprobe bestand insgesamt aus 1 905 147 SchülerInnen im Alter von 15-16 Jahren. Die Daten wurden im Zuge der PISA Studie der letzten Jahre (2000, 2003, 2006, 2009, 2012) erhoben. Insgesamt hatten bis zum Jahr 2012 79 Länder bzw. Länderregionen zu mindestens einem Zeitpunkt an der PISA Studie teilgenommen. Die genaue Anzahl an teilnehmenden SchülerInnen pro Jahr und Land kann Tabelle 1 entnommen werden.

Tabelle 1

Ursprungsstichprobe – Anzahl der teilnehmenden SchülerInnen der gesamten PISA Studie bis 2012 gegliedert nach Land und Jahr

Land	2000	2003	2006	2009	2012	Total
Albanien	4980	0	0	4596	4743	**14319**
Argentinien	3983	0	4339	4774	5908	**19004**
Australien	5176	12551	14170	14251	14481	**60629**
Austria	4745	4597	4927	6590	4755	**25614**
Azerbaijan	0	0	5184	4691	0	**9875**
Belgien	6670	8796	8857	8501	8597	**41421**
Brasilien	4893	4452	9295	20127	19204	**57971**
Bulgarien	4657	0	4498	4507	5282	**18944**
Kanada	29687	27953	22646	23207	21544	**125037**
Chile	4889	0	5233	5669	6856	**22647**
China-Taipei	0	0	8815	5831	6046	**20692**
Columbien	0	0	4478	7921	9073	**21472**
Connecticut (USA)	0	0	0	0	1697	**1697**
Costa Rica	0	0	0	4578	4602	**9180**
Kroatien	0	0	5213	4994	5008	**15215**
Tschechische RepubliK	5365	6320	5932	6064	5327	**29008**
Dänemark	4235	4218	4532	5924	7481	**26390**
Estland	0	0	4865	4727	4779	**14371**
Finland	4864	5796	4714	5810	8829	**30013**
Florida (USA)	0	0	0	0	1896	**1896**
Frankreich	4673	4300	4716	4298	4613	**22600**
Georgien	0	0	0	4646	0	**4646**
Deutschland	5073	4660	4891	4979	5001	**24604**
Griechenland	4672	4627	4873	4969	5125	**24266**
Hong Kong-China	4405	4478	4645	4837	4670	**23035**
Ungarn	4887	4765	4490	4605	4810	**23557**
Island	3372	3350	3789	3646	3508	**17665**
Indien	0	0	0	4826	0	**4826**
Indonesien	7368	10761	10647	5136	5622	**39534**
Irland	3854	3880	4585	3937	5016	**21272**
Israel	4498	0	4584	5761	5055	**19898**
Italien	4984	11639	21773	30905	31073	**100374**
Japan	5256	4707	5952	6088	6351	**28354**
Jordanien	0	0	6509	6486	7038	**20033**

Kasachstan	0	0	0	5412	5808	**11220**
Korea	4982	5444	5176	4989	5033	**25624**
Kirgisistan	0	0	5904	4986	0	**10890**
Lettland	3893	4627	4719	4502	4306	**22047**
Liechtenstein	314	332	339	329	293	**1607**
Litauen	0	0	4744	4528	4618	**13890**
Luxembourg	3528	3923	4567	4622	5258	**21898**
Macao-China	0	1250	4760	5952	5335	**17297**
Mazedonien	4510	0	0	0	0	**4510**
Malaysia	0	0	0	4999	5197	**10196**
Malta	0	0	0	3453	0	**3453**
Massachusetts (USA)	0	0	0	0	1723	**1723**
Mauritius	0	0	0	4654	0	**4654**
Mexiko	4600	29983	30971	38250	33806	**137610**
Miranda-Venezuela	0	0	0	2901	0	**2901**
Montenegro	0	0	4455	4825	4744	**14024**
Niederlande	2503	3992	4871	4760	4460	**20586**
Neuseeland	3667	4511	4823	4643	4291	**21935**
Norwegen	4147	4064	4692	4660	4686	**22249**
Panama	0	0	0	3969	0	**3969**
Peru	4429	0	0	5985	6035	**16449**
Polen	3654	4383	5547	4917	4607	**23108**
Portugal	4585	4608	5109	6298	5722	**26322**
Qatar	0	0	6265	9078	10966	**26309**
Moldawien	0	0	0	5194	0	**5194**
Rumänien	4829	0	5118	4776	5074	**19797**
Russland	6701	5974	5799	5308	5231	**29013**
Serbien	0	0	4798	5523	4684	**15005**
Shanghai-China	0	0	0	5115	5177	**10292**
Singapur	0	0	0	5283	5546	**10829**
Slowakei	0	7346	4731	4555	4678	**21310**
Slowenien	0	0	6595	6155	5911	**18661**
Spanien	6214	10791	19604	25887	25313	**87809**
Schweden	4416	4624	4443	4567	4736	**22786**
Schweiz	6100	8420	12192	11812	11229	**49753**
Thailand	5340	5236	6192	6225	6606	**29599**
Trinidad und Tobago	0	0	0	4778	0	**4778**
Tunesien	0	4721	4640	4955	4407	**18723**
Türkei	0	4855	4942	4996	4848	**19641**
Vereinigte Arabische Emirate	0	0	0	10867	11500	**22367**

Großbritannien	9340	9535	13152	12179	12659	**56865**
USA	3846	5456	5611	5233	4978	**25124**
Uruguay	0	5835	4839	5957	5315	**21946**
Jugoslawien	0	4405	0	0	0	**4405**

4.4.2. Arbeitsstichprobe

Die zur Beantwortung der Fragestellungen verwendete Stichprobe ergab sich durch die Einschränkung der Ursprungsstichprobe nach bereits genannten Kriterien (siehe Kapitel 4.2.). Sie bestand aus insgesamt 315 072 SchülerInnen aus 34 Ländern. Der folgenden Tabelle 2 kann die genaue Liste der Länder, die Gesamtanzahl der SchülerInnen pro Land sowie die genaue Geschlechterverteilung der SchülerInnen für jedes Jahr und Land entnommen werden.

Tabelle 2

Arbeitsstichprobe – Anzahl der SchülerInnen nach Einschränkung der Ursprungsstichprobe, gegliedert nach Land, Jahr und Geschlecht (m|w)

Land	2000 m|w	2003 m|w	2006 m|w	2009 m|w	Gesamt
Australien	1274 | 1358	1952 | 2183	2227 | 2425	2157 | 2476	**16052**
Austria	951 | 1113	479 | 647	576 | 694	626 | 898	**5984**
Belgien	1555 | 1633	1371 | 1428	1486 | 1504	1319 | 1353	**11649**
Brasilien	430 | 464	243 | 321	557 | 715	1891 | 2491	**7112**
Kanada	7288 | 7970	4219 | 4805	3415 | 3835	3356 | 3827	**38715**
Tschechische Republik	854 | 1163	820 | 931	778 | 901	705 | 833	**6985**
Dänemark	825 | 809	529 | 541	591 | 692	761 | 831	**5579**
Finnland	1169 | 1375	984 |1116	744 | 939	885 | 1128	**8340**
Frankreich	812 | 914	377 | 536	467 | 607	383 | 488	**4584**
Deutschland	901 | 1068	518 | 665	607 | 668	618 | 747	**5792**
Griechenland	808 | 869	389 | 488	494 | 631	612 | 706	**4997**
Hong Kong – China	992 | 1142	741 | 803	824 | 967	885 | 921	**7275**
Ungarn	791 | 865	700 | 658	564 | 683	599 | 705	**5565**
Island	747 | 851	484 | 543	482 | 587	513 | 605	**4812**
Indonesien	900 | 927	781 | 932	1127 | 1210	440 | 469	**6786**
Irland	944 | 1151	669 | 733	683 | 861	555 | 660	**6256**
Italien	789 | 878	1278 | 1620	2318 | 2766	3964 | 4468	**18081**
Japan	1020 | 980	511 | 616	713 | 742	793 | 821	**6196**
Korea	1551 | 1275	1146 | 832	952 | 1049	1007 | 990	**8802**
Lettland	577 | 714	509 | 729	613 | 850	586 | 775	**5353**
Liechtenstein	57 |65	56 | 50	37 | 73	44 | 54	**436**
Luxembourg	500 | 566	371 | 460	519 | 598	520 | 601	**4135**
Mexiko	864 | 874	3208 | 3901	4089 | 5120	5680 | 6565	**30301**
Niederlande	784 | 864	852 | 884	1059 | 1050	990 | 1039	**7522**
Neuseeland	897 | 1030	707 | 753	719 | 893	760 | 836	**6595**
Norwegen	960 | 931	447 | 516	601 | 620	631 | 746	**5452**
Polen	616 | 659	487 | 620	744 | 901	695 | 908	**5630**
Portugal	806 | 942	455 | 593	564 | 725	751 | 948	**5784**
Russland	1022 | 1204	567 | 663	638 | 691	564 | 694	**6043**
Spanien	1370 | 1396	1349 | 1649	2579 | 2868	3556 | 3518	**18285**
Schweden	1026 | 1025	599 | 670	560 | 646	576 | 637	**5739**
Schweiz	1195 | 1252	1066 | 1154	1641 | 1826	1550 | 1719	**11403**
Thailand	1062 | 1568	562 | 800	677 | 1071	800 | 1131	**7671**
Großbritannien	2359 | 2419	1420 | 1631	1751 | 2035	1709 | 1837	**15161**

5. Ergebnisse

In den folgenden Kapiteln werden die genauen Ergebnisse der Analysen dargestellt. Zuerst werden die Resultate der Berechnungen für jedes Jahr vorgestellt und danach für jedes einzelne Land, sowohl bezüglich des Teilungskriteriums Anzahl gelöster Aufgaben als auch Geschlecht.

5.1. Ergebnisse für jedes Jahr

Die Berechnung der Itemparameter sowie der LRTs wurde für jedes Jahr separat, über alle Länder hinweg, durchgeführt. In Tabelle 3 ist ersichtlich, dass sowohl in Bezug auf das Teilungskriterium Anzahl gelöster Aufgaben, als auch in Bezug auf das Teilungskriterium Geschlecht, in jedem Jahr ein signifikanter LRT von Andersen resultiert (jeweils $p < 0{,}0001$).

Aus diesem Grund wurde für jedes Jahr eine graphische Modellkontrolle durchgeführt. In der Tabelle 3 sind jene Items angeführt, deren Differenz der Itemparameter der zwei Teilstichproben mehr als ein Zehntel der Spannweite der Parameterschätzungen beträgt.
Es ist ersichtlich, dass bezüglich des Merkmals Anzahl gelöster Aufgaben in jedem Jahr mindestens zwei Items in den beiden Teilstichproben (Personen mit niedriger Anzahl gelöster Aufgaben vs. Personen mit hoher Anzahl gelöster Aufgaben) deutlich andere Schwierigkeiten aufweisen und somit nicht Rasch-Modell konform sind. Auch bezüglich des Merkmals Geschlecht sieht man, dass das Item mit der Nummer 13 in jedem Jahr für Männer und Frauen eine andere Schwierigkeit aufweist. Es fällt Schülerinnen unverhältnismäßig schwerer als Schülern (siehe Abbildungen 8-11 der graphischen Modellkontrollen im Anhang).

Somit kann nach inhaltlich relevanten Kriterien davon ausgegangen werden, dass nicht alle Items Rasch-Modell konform sind.

Tabelle 3

Ergebnisse des Andersen-Likelihood-Ratio-Tests sowie Items, die bei der graphischen Modellkontrolle nicht dem Rasch-Modell entsprechen, für jedes Jahr bezüglich Anzahl gelöster Aufgaben und Geschlecht

Jahr	N	Anzahl gelöster Aufgaben				Geschlecht			
		x^2 (LRT)	FG	p-Wert	Items	x^2 (LRT)	FG	p-Wert	Items
2000	81010	598.039	19	< 0,0001	i2, i5, i9,i16	907.311	19	< 0,0001	i13
2003	66317	943.995	19	< 0,0001	i5, i16	852.901	19	< 0,0001	i13
2006	78839	1707.532	19	< 0,0001	i10, i16, i18	1063.484	19	< 0,0001	i13
2009	88906	1861.511	19	< 0,0001	i10, i16, i18	1143.898	19	< 0,0001	i13

Abbildung 2 soll exemplarisch für die graphischen Modellkontrollen angeführt werden. Liegen die Itempunkte innerhalb der strichlierten Linie, beträgt die Differenz der Itemparameter der zwei Teilstichproben weniger als ein Zehntel der Spannweite der Parameterschätzungen und somit kann davon ausgegangen werden, dass diese Items dem Rasch-Modell entsprechen. Alle weiteren Abbildungen der graphischen Modellkontrollen für die restlichen Jahre, sowohl für das Teilungskriterium Anzahl gelöster Aufgaben, als auch für Geschlecht, sind im Anhang zu finden.

In Abbildung 2 ist ersichtlich, dass im Jahr 2006 Personen mit niedriger Anzahl gelöster Aufgaben (<= Median) die Items mit den Nummen 16 und 18, im Vergleich zu den übrigen Aufgaben, unverhältnismäßig schwerer fällt als dies für Personen mit hoher Anzahl (> Median) gelöster Aufgaben der Fall ist; bei dem Item 10 ist es umgekehrt.

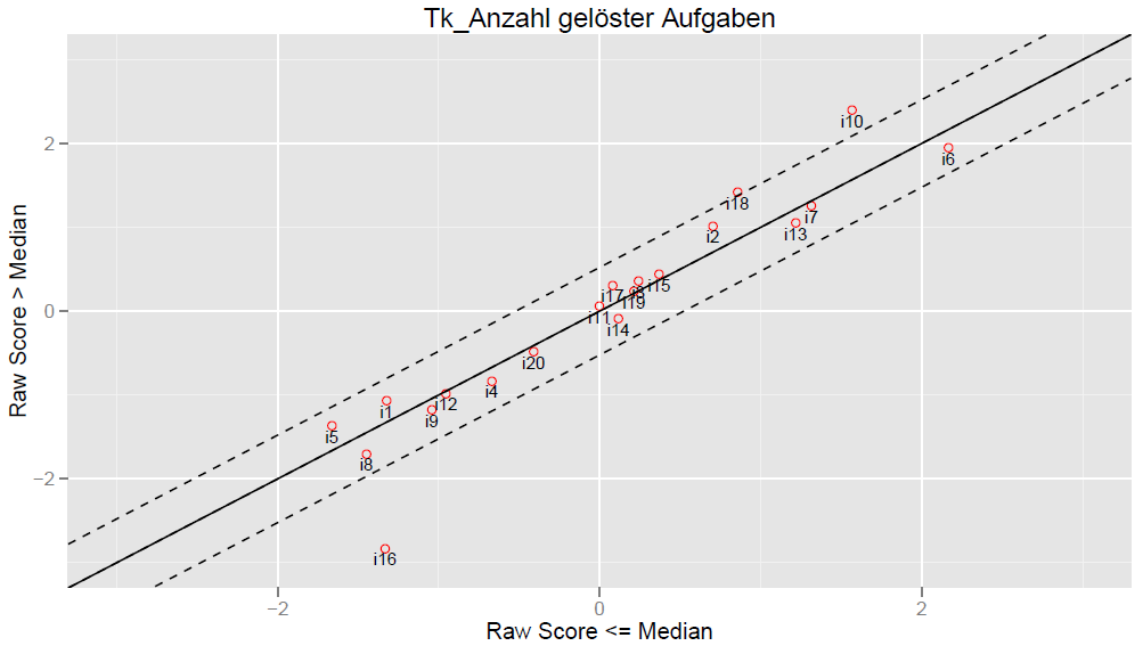


Abbildung 2. Graphische Modellkontrolle für das Jahr 2006 in Bezug auf das Teilungskriterium *Anzahl gelöster Aufgaben*

5.2. Ergebnisse für jedes Land

Die Berechnung der Itemparameter sowie der LRTs wurde für jedes Land separat, über alle Jahre hinweg, durchgeführt. Als Teilungskriterien wurden die Anzahl gelöster Aufgaben sowie Geschlecht herangezogen.

5.2.1. Teilungskriterium Anzahl gelöster Aufgaben

In Tabelle 4 sind die Ergebnisse des LRT für jedes Land, bezüglich des Teilungskriteriums Anzahl gelöster Aufgaben, dargestellt. In den Ländern Finnland (i16), Deutschland (i16), Hong Kong (i16), Island (i16), Irland (i8), Liechtenstein (i4, i5 und i16), Neuseeland (i16) und Schweden (i12 und i16) mussten, auf Grund ungeeigneter Antwortmuster, Items aus der Analyse ausgeschlossen werden. In jedem Land, außer Liechtenstein (x^2= 10,008; df=16;

p= 0,866), resultierte ein signifikanter LRT nach Andersen (jeweils p < 0,0001). Somit wurde für jedes Land, außer Liechtenstein, eine graphische Modellkontrolle durchgeführt. Wie in der letzten Spalte der Tabelle 4 ersichtlich ist, ist in jedem Land mindestens ein Item bei der graphischen Modellkontrolle auffällig und entspricht nicht dem Rasch-Modell. Die Abbildungen der graphischen Modellkontrollen für alle Länder sind im Anhang angeführt (Abbildungen 12-44).

Es kann nach praktisch relevanten Kriterien davon ausgegangen werden, dass nicht alle Items Rasch-Modell konform sind. Nur in Liechtenstein entsprechen die Items, laut dem LRT, dem Rasch-Modell.

Tabelle 4

Ergebnisse des Andersen-Likelihood-Ratio-Tests sowie Items, die bei der graphischen Modellkontrolle nicht dem Rasch-Modell entsprechen, für jedes Land betreffs Anzahl gelöster Aufgaben

Land	N	x^2 (LRT)	FG	p-Wert	Items
Australien	16052	217,950	19	< 0,0001	i8, i10, i12, i16
Austria	5984	63,622	19	< 0,0001	i8, i9, i12, i18, i19
Belgien	11649	133,787	19	< 0,0001	i16
Brasilien	7112	261,479	19	< 0,0001	i1,i5,i8,i16,i18
Kanada	38715	426,928	19	< 0,0001	i1, i8, i10, i16
Tschechische Republik	6985	82,657	19	< 0,0001	i4, i10, i16
Dänemark	5579	59,293	19	< 0,0001	i16, i20
Finnland	8340	59,774	18	< 0,0001	i8
Frankreich	4584	66,677	19	< 0,0001	i10, i15, i16, i18
Deutschland	5792	74,656	18	< 0,0001	i10, i15, i17
Griechenland	4997	131,482	19	< 0,0001	i5, i8, i10, i16
Hong Kong – China	7275	76,766	18	< 0,0001	i5
Ungarn	5565	119,547	19	< 0,0001	i8, i10, i16
Island	4812	106,830	18	< 0,0001	i1, i8, i10, i18
Indonesien	6786	247,534	19	< 0,0001	i10, i16, i18
Irland	6256	74,194	18	< 0,0001	i5, i10, i11, i12, i16
Italien	18081	257,020	19	< 0,0001	i5, i10, i16,i18
Japan	6196	59,417	19	< 0,0001	i5, i16
Korea	8802	70,353	19	< 0,0001	i5
Lettland	5353	192,094	19	< 0,0001	i8, i10, i16
Liechtenstein	436	10,008	16	0,866	–
Luxembourg	4135	55,517	19	< 0,0001	i1, i10, i16
Mexiko	30301	1179,012	19	< 0,0001	i10, i16
Niederlande	7522	143,108	19	< 0,0001	i10, i16
Neuseeland	6595	56,244	18	< 0,0001	i8, i10, i12
Norwegen	5452	79,726	19	< 0,0001	i10, i16
Polen	5630	87,735	19	< 0,0001	i8, i10, i16, i18
Portugal	5784	98,428	19	< 0,0001	i1, i10, i16
Russland	6043	104,103	19	< 0,0001	i5, i16
Spanien	18285	487,194	19	< 0,0001	i16, i18
Schweden	5739	67,188	17	< 0,0001	i4, i8, i10, i18, i20
Schweiz	11403	133,981	19	< 0,0001	i10, i16, i18
Thailand	7671	492,722	19	< 0,0001	i5, i10, i16
Großbritannien	15161	197,323	19	< 0,0001	i8, i12, i16

5.2.2. Teilungskriterium Geschlecht

In Tabelle 5 sind die Ergebnisse des LRT für jedes Land, bezüglich des Teilungskriteriums Geschlecht, dargestellt. Auf Grund des ungeeigneten Antwortmusters, musste in Liechtenstein ein Item (i16) aus der Analyse ausgeschlossen werden. Im LRT resultieren, mit Ausnahme von Liechtenstein (x^2= 19.238; df= 18; p= 0,377), durchwegs signifikante Ergebnisse; abgesehen von Frankreich (x^2= 44.759; df= 19; p= 0,001) resultiert in jedem Land ein p-Wert < 0,0001. Daher wurde für jedes Land, außer Liechtenstein, eine graphische Modellkontrolle durchgeführt.

In den Ländern Indonesien, Luxembourg und Spanien ergab sich keine Differenz der Itemparameter der zwei Teilstichproben (Männer vs. Frauen) um mehr als ein Zehntel der Spannweite der Parameterschätzungen. Somit kann, trotz statistisch signifikantem Ergebnis im LRT, nach praktisch relevanten Kriterien davon ausgegangen werden, dass die Items in diesen Ländern dem Rasch-Modell entsprechen. Auch für Liechtenstein kann von einer Modellgültigkeit ausgegangen werden.

Wie in der letzten Spalte der Tabelle 5 ersichtlich ist, ist in den übrigen Ländern mindestens ein Item bei der graphischen Modellkontrolle auffällig und entspricht nicht dem Rasch-Modell. Hierbei fällt besonders das Item mit der Nummer 13 auf, das in allen Ländern, außer Island und Norwegen, bei Männern und Frauen unterschiedliche Schwierigkeiten aufweist.

Tabelle 5

Ergebnisse des Andersen-Likelihood-Ratio-Tests sowie Items, die bei der graphischen Modellkontrolle nicht dem Rasch-Modell entsprechen, für jedes Land bezüglich des Teilungskriteriums Geschlecht

Land	N	x^2 (LRT)	FG	p-Wert	Items
Australien	16052	189,955	19	< 0,0001	i13
Austria	5984	99,901	19	< 0,0001	i13, i19
Belgien	11649	208,144	19	< 0,0001	i13
Brasilien	7112	107,501	19	< 0,0001	i13
Kanada	38715	456,403	19	< 0,0001	i13
Tschechische Republik	6985	175,600	19	< 0,0001	i13
Dänemark	5579	92,447	19	< 0,0001	i13
Finnland	8340	134,936	19	< 0,0001	i13
Frankreich	4584	44,759	19	0,001	i13
Deutschland	5792	124,349	19	< 0,0001	i13
Griechenland	4997	136,947	19	< 0,0001	i11, i13
Hong Kong – China	7275	185,858	19	< 0,0001	i11, I13
Ungarn	5565	150,624	19	< 0,0001	i1, i19, i13
Island	4812	89,588	19	< 0,0001	i6, i12
Indonesien	6786	83,848	19	< 0,0001	–
Irland	6256	85,648	19	< 0,0001	i13
Italien	18081	411,603	19	< 0,0001	i13
Japan	6196	93,304	19	< 0,0001	i13, i16
Korea	8802	207,584	19	< 0,0001	i13
Lettland	5353	121,353	19	< 0,0001	i13
Liechtenstein	436	19,238	18	0,377	–
Luxembourg	4135	48,630	19	< 0,0001	–
Mexiko	30301	394,132	19	< 0,0001	i13
Niederlande	7522	172,428	19	< 0,0001	i8, i13
Neuseeland	6595	81,990	19	< 0,0001	i13
Norwegen	5452	95,964	19	< 0,0001	i16
Polen	5630	118,593	19	< 0,0001	i1, i13
Portugal	5784	105,465	19	< 0,0001	i1, i13
Russland	6043	212,490	19	< 0,0001	i5, i13
Spanien	18285	336,309	19	< 0,0001	–
Schweden	5739	94,734	19	< 0,0001	i8, i13, i16
Schweiz	11403	150,340	19	< 0,0001	i13
Thailand	7671	197,200	19	< 0,0001	i13
Großbritannien	15161	220,972	19	< 0,0001	i13

Auf der folgenden Abbildung 3 erkennt man, dass das Item mit der Nummer 13 in Belgien für Männer und Frauen eine andere Schwierigkeit aufweist. Es fällt Schülerinnen unverhältnismäßig schwerer als männlichen Schülern. Die Abbildungen der graphischen Modellkontrollen für die restlichen Länder, für das Teilungskriterium Geschlecht, können dem Anhang entnommen werden (Abbildungen 45-77).

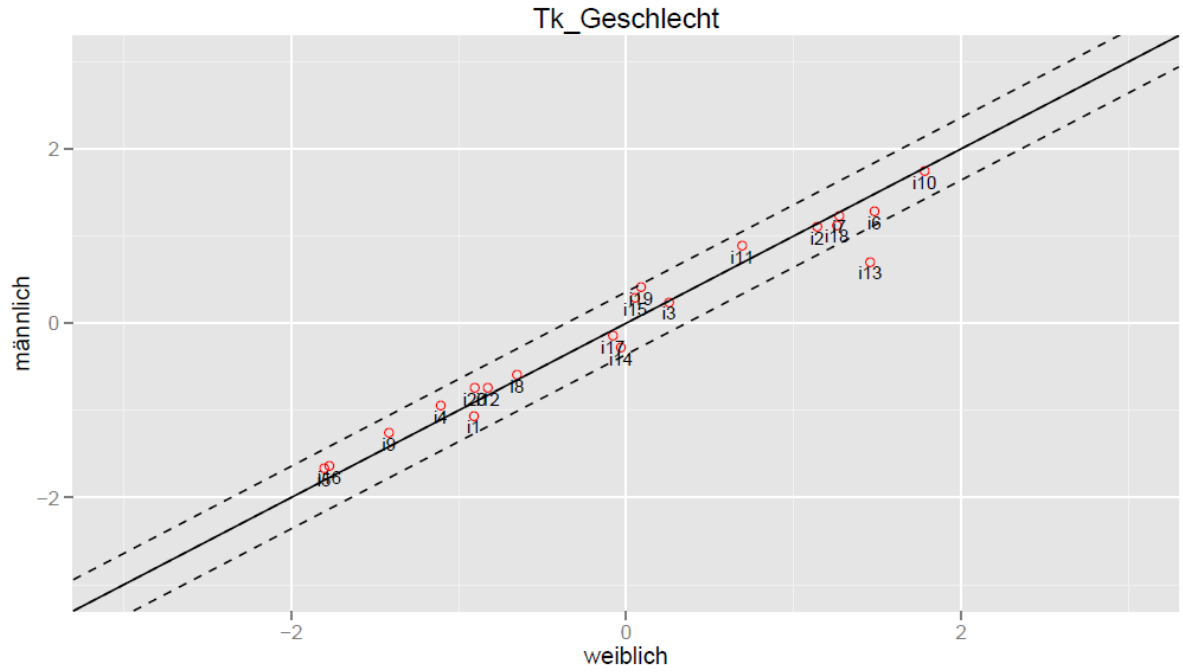


Abbildung 3. Graphische Modellkontrolle für das Land Belgien in Bezug auf das Teilungskriterium *Geschlecht*.

6. Diskussion

Die Analyse des Datensatzes ergibt, dass über 90% aller durchgeführten Modellüberprüfungen nach praktisch relevanten Kriterien gegen die Gültigkeit des Rasch-Modells sprechen.

In jedem Jahr der PISA Studie waren im Kompetenzbereich Lesen in Bezug auf Anzahl gelöster Aufgaben mindestens zwei Items bezüglich dem praktisch relevanten Kriterium nicht Rasch-Modell konform. Hierbei ist besonders das Item mit der Nummer 16 hervorzuheben, das in jedem Jahr auffällig ist. Dieses Item fällt SchülerInnen mit niedriger Anzahl gelöster Aufgaben schwerer, als SchülerInnen mit hoher Anzahl gelöster Aufgaben. Somit kann nicht davon ausgegangen werden, dass alle Items Rasch-Modell konform sind.
Dieses Ergebnis entspricht jenem von Allerup (2007), der zu dem Schluss kam, dass sich die Schwierigkeiten der Linking Items für die Jahre 2000 und 2003 unterscheiden. Auch Wetzel und Carstensen (2013) konnten einen Unterschied in den Itemschwierigkeiten zwischen den Jahren 2000 und 2009 nachweisen. Als mögliche Erklärung führten sie Variationen in der Itemformulierung zu den unterschiedlichen Zeitpunkten an. Ob das für die vorliegenden Daten in gleicher Weise als Begründung zutreffend ist, kann nicht mit Gewissheit gesagt werden, da die ausformulierten Items nicht vorliegen. Die Modifizierung von Linking Items in einzelnen Ländern stellt jedoch ein bekanntes Problem dar (OECD, 2014b), daher ist zu vermuten, dass die Modifizierung von Items auch Einfluss auf das vorliegende Ergebnis hatte.
Bezogen auf die Länder gibt es – mit Ausnahme von Liechtenstein – jeweils mindestens ein Item in Bezug auf das Teilungskriterium Anzahl gelöster Aufgaben, das gegen die Gültigkeit des Rasch-Modells spricht. Auch hier fällt das Item mit der Nummer 16 auf, da es in dreiviertel aller Länder nicht Rasch-Modell konform ist oder auf Grund von unpassenden Antwortmustern, von der Analyse ausgeschlossen werden musste. Abgesehen von Item Nummer 16 ergeben die anderen auffälligen Items kein einheitliches Bild. Bis zu fünf auffällige, nicht konforme Items finden sich in jedem Land, das entspricht einem Viertel aller überprüften Items. Daher kann davon ausgegangen werden, dass ein gültiger

Vergleich der Ergebnisse der PISA Studie zwischen den Ländern nicht möglich ist. Somit konnten die Studienergebnisse von Kreiner und Christensen (2014) auch nach Anwendung des praktisch relevanten Kriteriums bestätigt werden.

Bezogen auf das Teilungskriterium Geschlecht ist es Item Nummer 13, das jedes Jahr keine Rasch-Modell-Konformität nach dem praktisch relevanten Kriterium aufweist. Dieses Item fällt Schülerinnen deutlicher schwerer als Schülern. Auch Allerup (2007) konnte für das Jahr 2003 unterschiedliche Itemschwierigkeiten für Schüler und Schülerinnen nachweisen.
Bei Betrachtung der einzelnen Länder zeigt sich, dass auch hier Item Nummer 13 in 85% aller durgeführten graphischen Modellkontrollen eine praktisch relevante Abweichung aufweist. Ausnahmen stellen jene Länder dar, bei denen keine Items auffällig waren – Indonesien, Luxembourg und Spanien – sowie Island und Norwegen.
Da die vorliegende Studie nur mit einem kleinen Teil aller möglichen Items der PISA Studie durchgeführt wurde, ist anzunehmen, dass die Heranziehung anderer Items wahrscheinlich auch zu anderen Ergebnissen führen wird. Die vorliegenden Ergebnisse lassen dennoch den Schluß zu, dass die Items der PISA Studie nicht Rasch-Modell konform sind, da eine Abweichung vom Rasch-Modell bei einem Subset von Items gleichzeitig eine Modellungültigkeit bei dem kompletten Itemset impliziert (Kreiner & Christensen, 2014). Die vorliegenden Hinweise auf DIF könnten die in der PISA Studie in Bezug auf Geschlecht berichteten Ergebnisse im Kompetenzbereich Lesen (2014a) beeinflussen. Da bei einem Test, der DIF aufweist, gleiche Testwerte nicht mehr gleiche Leistungen wiederspiegeln (Kubinger, 2015).
Allerdings sind – abgesehen von dem Item mit der Nummer 13 – nur vereinzelt andere Items auffällig, weshalb zu vermuten ist, dass bereits nach dem Ausschluss dieses Items im Großteil aller Länder die übrigen analysierten Items Rasch-Modell konform wären. Doch diese Hypothese müsste erst mittels neuer Berechnungen überprüft werden.

Die Anwendung dieses praktisch relevanten Kriteriums – nämlich die Differenz der Itemparameterschätzungen der zwei Teilstichproben um mehr als ein Zehntel der Spannweite – stellt jedoch nur eine Möglichkeit dar, um die Items bezüglich ihrer Rasch-Modell-Konformität zu überprüfen. Sie wird als Faustregel

herangezogen (Kubinger, 2005), ist jedoch nicht explizit auf die vorliegende Fragestellung zugeschnitten. In zukünftigen Studien sollte ein spezifischeres Kriterium zur Überprüfung der Rasch-Modellgültigkeit herangezogen werden. Hierzu könnten kritische Itemparameterdifferenzen ermittelt werden, die zu einer praktisch relevanten Änderung im Personenfähigkeitsparameter führen (Kubinger, 2005). Zu diesem Zweck könnte eine Simulationsstudie durchgeführt werden (vgl. Kubinger, Rasch & Yanagida, 2009). Dabei werden Items simuliert, von denen sich ein Itempaar bzw. mehrere Itempaare in der Schwierigkeit in zwei Teilgruppen unterscheiden, also DIF aufweisen. Dadurch kann der Einfluss von DIF auf die Personenfähigkeit ermittelt werden. Daraus kann abgeleitet werden, wie groß die Differenz der Itemparameterschätzungen in zwei Teilstichproben exakt sein muss, damit es zu einer praktisch relevanten Änderung in den Personenfähigkeiten kommt.

7. Zusammenfassung

Diese Arbeit setzte sich zum Ziel die Items der PISA Studie bezüglich ihrer Rasch-Modell-Konformität zu überprüfen.

Die PISA Studie basiert auf dem dichotom logistischen Modell von Rasch (1980) und wurde zum ersten Mal im Jahr 2000 von den Mitgliedsstaaten der OECD (englisch für: Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) durchgeführt und seit dem alle drei Jahre wiederholt. Ihr Ziel ist es die Kompetenzen von 15 jährigen SchülerInnen in den Bereichen Lesen, Mathematik und Naturwissenschaften zu erfassen, um den teilnehmenden Staaten vergleichbare Daten zum Stand ihrer Bildungssysteme zur Verfügung zu stellen (OECD, 2012). Die verwendeten Items müssen bestimmte Vorraussetzungen erfüllen, um dem Rasch-Modell zu entsprechen und somit einen gültigen Vergleich der Ergebnisse von unterschiedlichen Ländern zu gewährleisten (Fischer, 1974).
Ergebnisse bereits durchgeführter Studien weisen darauf hin, dass die Items der PISA Studie die Vorraussetzungen des Rasch-Modells nicht erfüllen (z.B. Kreiner und Christensen, 2014). Beispielsweise ergaben die Analysen von Wetzel und Carstensen (2013) Differenzen in den Itemschwiergkeiten zwischen unterschiedlichen Zeitpunkten und Allerup (2007) fand Unterschiede in den Itemschwierigkeiten für Schülerinnen und männliche Schüler.

Daher wurde in dieser Forschungsarbeit untersucht, ob die Items der PISA Studie Rasch-Modell konform sind und zwar in Bezug auf die Länder, das Geschlecht und die unterschiedlichen Zeitpunkte. Analysiert wurden 20 Linking Items der PISA Studie – also jene Items, die in jedem Jahr vorgegeben wurden – aus den Jahren 2000, 2003, 2006 und 2009 aus dem Kompetenzbereich Lesen von 34 Ländern, die zu allen Zeitpunkten an der Studie teilgenommen haben. Insgesamt wurde ein Datensatz von 34 Ländern, 20 Items und 315 072 SchülerInnen ausgewertet. Die Auswertung wurde mittels Andersen-Likelihood-Ratio-Test (LRT; Andersen, 1973) für jedes Jahr und für jedes Land mit dem Teilungskriterium *Anzahl gelöster Aufgaben* und *Geschlecht* durchgeführt. Für

alle signifikanten Ergebnisse des LRT wurden graphische Modellkontrollen durchgeführt (Fischer, 1974). Da es wahrscheinlich war, bereits auf Grund der ungewöhnlich großen Stichprobe, signifikante Ergebnisse zu erhalten (Kubinger, 2005), wurden für diese Arbeit nicht signifikante, sondern praktisch relevante Abweichungen als Kriterium für die Modellgültigkeit gewählt. Eine Abweichung von der 45°-Geraden in der graphischen Modellkontrolle wurde als praktisch relevant angesehen, wenn die Differenz der Itemparameterschätzungen aus zwei Teilstichproben mehr als ein Zehntel der Spannweite der Parameterschätzungen betrug (Goethals, 1994).

Die Analyse des Datensatzes ergab, dass in jedem Jahr der PISA Studie sowie in jedem Land im Kompetenzbereich Lesen sowohl in Bezug auf Anzahl gelöster Aufgaben als auch Geschlecht mindestens ein Item bezüglich dem praktisch relevanten Kriterium, nicht Rasch-Modell konform ist.
Bezogen auf die Länder gibt es – mit Ausnahme von Lichtenstein – jeweils mindestens ein Item in Bezug auf das Teilungskriterium Anzahl gelöster Aufgaben, das nicht Rasch-Modell konform ist. Nur bezüglich des Teilungskriteriums Geschlecht wurde in drei Ländern keine praktisch relevanten Abweichungen gefunden, und zwar in Indonesien, Luxembourg und Spanien. Das bedeutet, dass beinahe in jedem Land und in Bezug auf beide Teilungskriterien Items vorhanden sind, die in den Teilgruppen nicht dieselbe Schwierigkeit aufweisen. Somit kann nicht davon ausgegangen werden, dass die Items der PISA Studie im Kompetenzbereich Lesen Rasch-Modell konform sind. Die Ergebnisse bestätigen jene aus bereits durchgeführten Studien (Allerup, 2007; Kreiner und Christensen,2014; Wetzel und Carstensen, 2013). Eine mögliche Erklärung stellen Variationen in der Itemformulierung zu den unterschiedlichen Zeitpunkten dar (OECD, 2014b).
Da in der vorliegenden Arbeit auffällt, dass für das Teilungskriterium Anzahl gelöster Aufgaben bis zu einem Viertel der überprüften Items nicht Rasch-Modell konform sind, aber für das Teilungskriterium Geschlecht nur primär ein Item gegen die Annahme der Gültigkeit des Rasch-Modells spricht, wäre es interessant zu überprüfen, ob nach Ausschluss dieses Items mehr Länder dem Rasch-Modell entsprechen würden. Es wäre auch sinnvoll ein spezifischeres Kriterium zur Überprüfung der Rasch-Modellgültigkeit heranzuziehen. Hierzu könnten

kritische Itemparameterdifferenzen ermittelt werden, die zu einer praktisch relevanten Änderung im Personenfähigkeitsparameter führen (Kubinger, 2005). Diese kritischen Differenzen könnten zum Beispiel mit Hilfe einer Simulationsstudie ermittelt werden (Kubinger et al., 2009).

8. Literatur

Adams, R. (2011). *Comments on Kreiner 2011: Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment*. Abgerufen von http://www.oecd.org/pisa/47681954.pdf

Allerup, P. (2007). Identification of Group Differences Using PISA Scales Considering Effects of Inhomogeneous Items. In S. T. Hopmann, G. Brinek & M. Retzl (Hrsg.) *PISA According to PISA* (S. 175-203). Wien: LIT Verlag.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika,* 38(1): 123-140.

Deutsches PISA-Konsortium (Hrsg.). (2003). *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–323). Opladen: Leske+Budrich.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of Education Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Goethals, R. (1994). *Die praktische Erprobung von Alternativen zur multiple choice Vorgabe bei Computertests* (Nicht veröffentlichte Dissertation). Universität Wien, Österreich.

Grossmann, W., Ledl, T., Neuwirth, E., Ponocny, I., & Steiner, P. (2006). Methodisch-statistische Grundlagen von PISA. In Neuwirth, E., Ponocny, I., & Grossmann, W. (Hrsg.), *PISA 2000 und PISA 2003: Vertiefende Analysen und Beiträge zur Methodik*. (S.11-134). Graz: Leykam.

Kreiner, S., & Christensen, K. B. (2014) Analyses of model fit and robustness, a new look at the Pisa scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210-231.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5(4)*, 377-394.

Kubinger, K. D. (2009). *Psychologische Diagnostik. Theorie und Praxis*

psychologischen Diagnostizierens. Göttingen: Hogrefe Verlag GmbH & Co. KG.

Kubinger, K. D. (2015). Rasch-Modell. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. Abgerufen von https://portal.hogrefe.com/dorsch/rasch-modell/

Kubinger, K. D., Rasch, D., & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly, 51(4)*, 370-384.

Mair, P., Hatzinger, R., & Maier, M. (2015). *eRm: Extended Rasch Modeling*. http://CRAN.R-project.org/package=eRm

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World* (Vol. 1). PISA, OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. PISA, OECD Publishing. *http://dx.doi.org/10.1787/9789264167872-en*

OECD (2014a). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014). PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD (2014b). *PISA 2012 Technical Report*. PISA, OECD Publishing.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna*:* R Foundation for Statistical Computing. http://www.R-project.org/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Schwantner, U., & Schreiner, C. (Hrsg.). (2010). *PISA 2009. Internationaler Vergleich von Schülerleistungen. Technischer Bericht*. Abgerufen von https://www.bifie.at/buch/1293

Wetzel, E., & Carstensen, C. H. (2013). Linking PISA 2000 and PISA 2009: Implications of instrument design on measure invariance. *Psychological Test and Assessment Modeling, 55 (2)*, 181-206.

Wu, M. L., Adams, R. J.,Wilson, M., & Haldane, S. (2007). *ConQuest Version 2.0. [Computer software]*. Assessment Systems Corporation, St. Paul, MN.

9. Anhang

9.1. Abbildungen der Graphischen Modellkontrollen

Im Folgenden finden sich die Abbildungen der graphischen Modellkontrollen der einzelnen Jahre in Bezug auf das Teilungskriterium *Anzahl gelöster Aufgaben*.

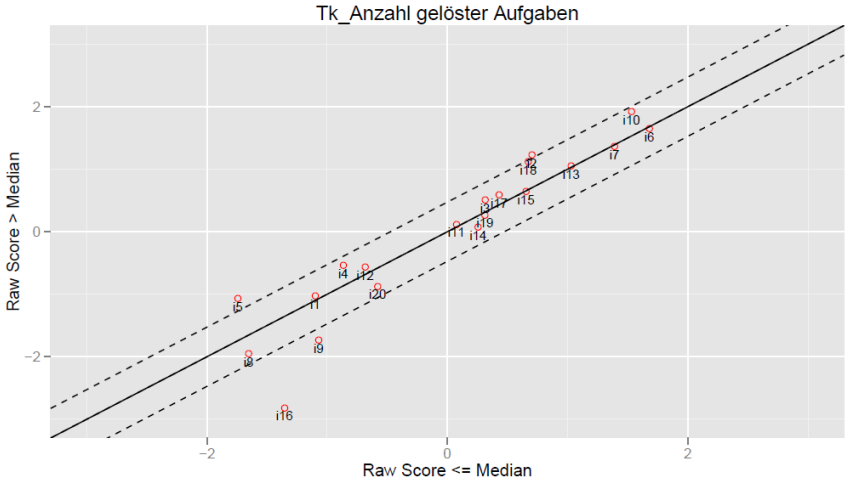


Abbildung 4. Jahr 2000

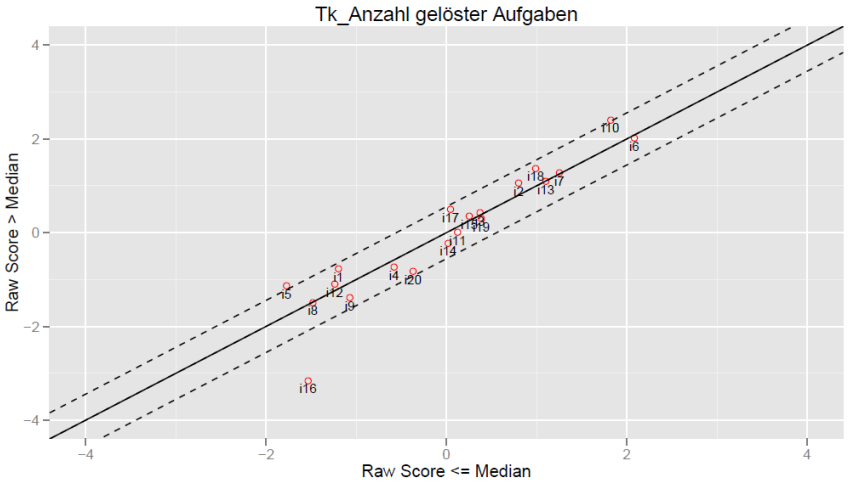


Abbildung 5. Jahr 2003

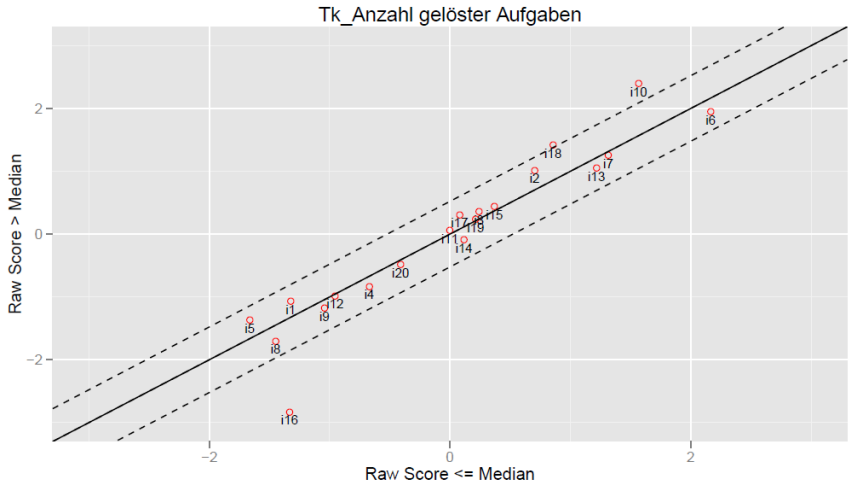


Abbildung 6. Jahr 2006

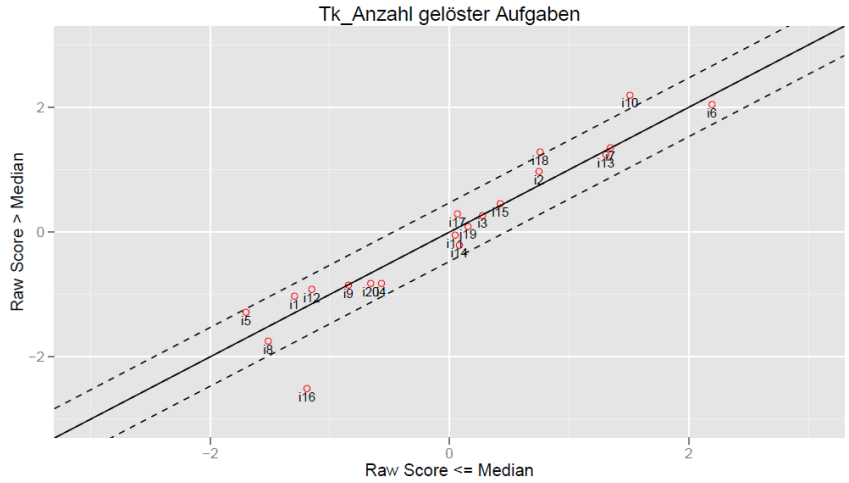


Abbildung 7. Jahr 2009

Im Folgenden finden sich die Abbildungen der graphischen Modellkontrollen der einzelnen Jahre in Bezug auf das Teilungskriterium Geschlecht.

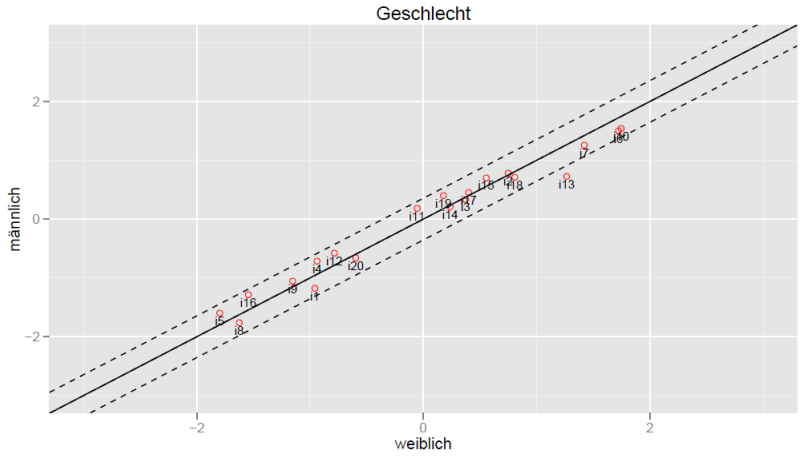


Abbildung 8. Jahr 2000

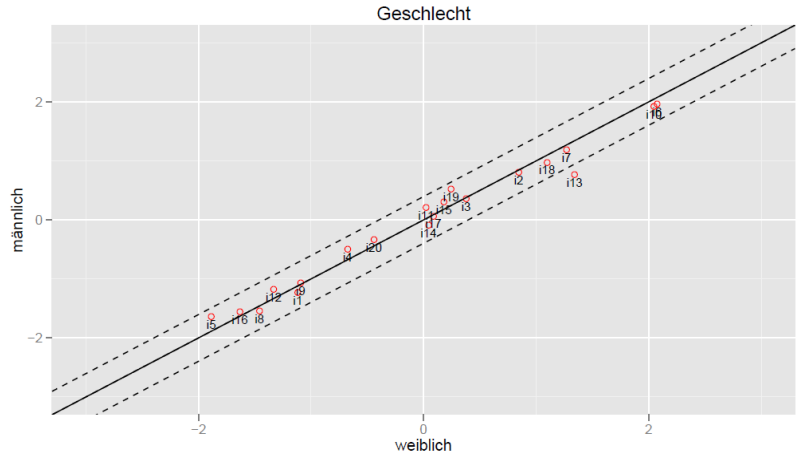


Abbildung 9. Jahr 2003

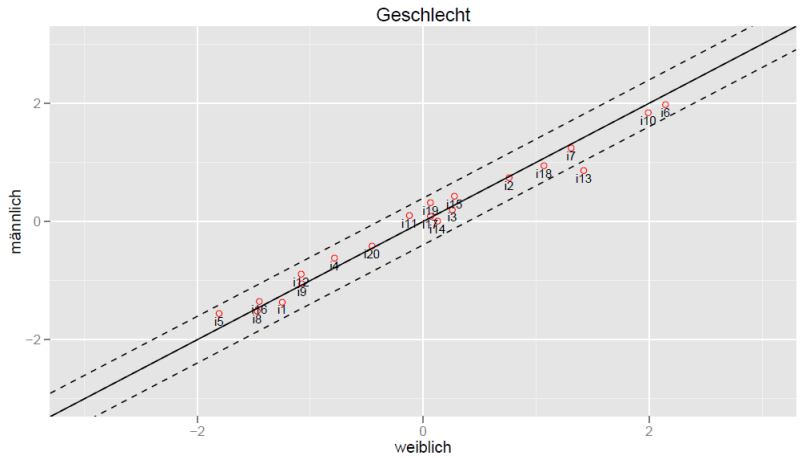


Abbildung 10. Jahr 2006

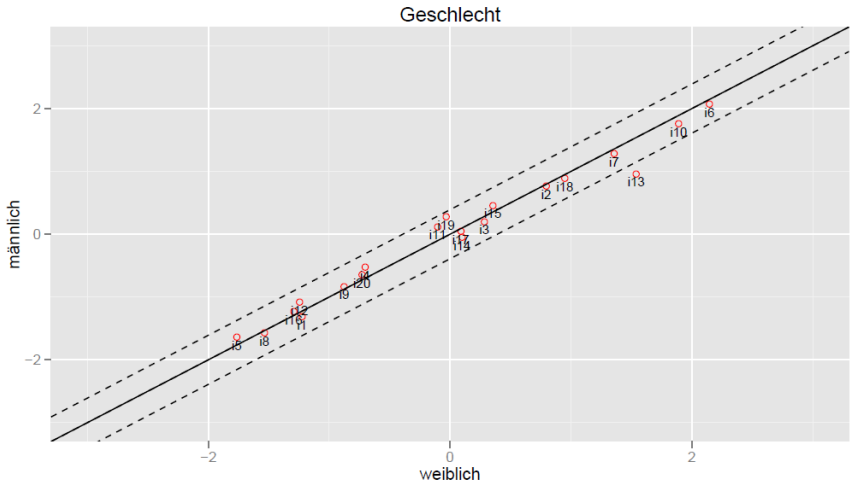


Abbildung 11. Jahr 2009

Im Folgenden finden sich die Abbildungen der graphischen Modellkontrollen der einzelnen Länder in Bezug auf das Teilungskriterium Anzahl gelöster Aufgaben.

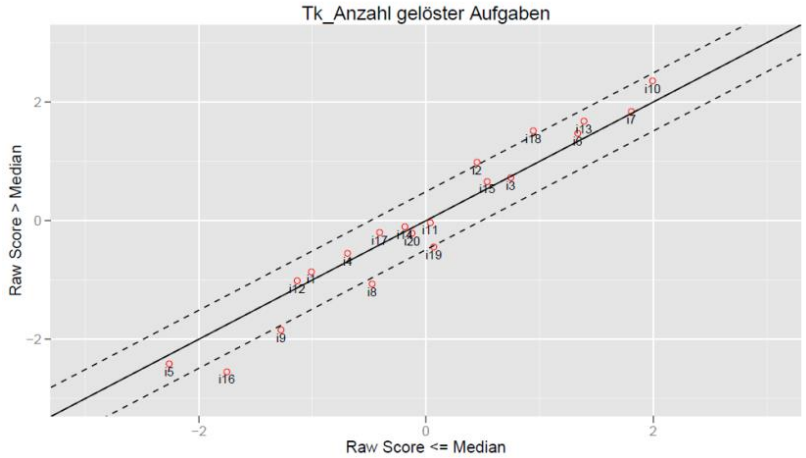


Abbildung 12. Österreich

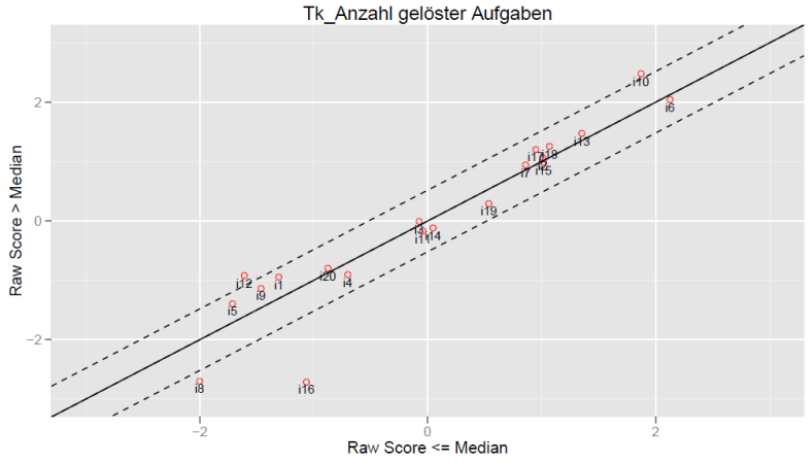


Abbildung 13. Australien

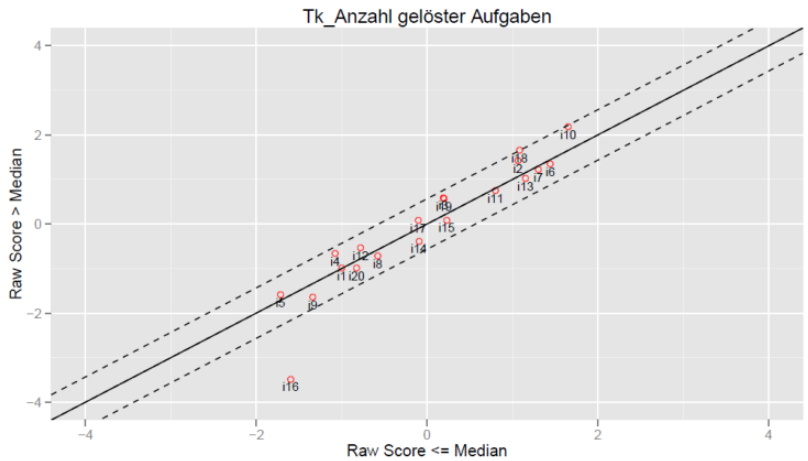


Abbildung 14. Belgien

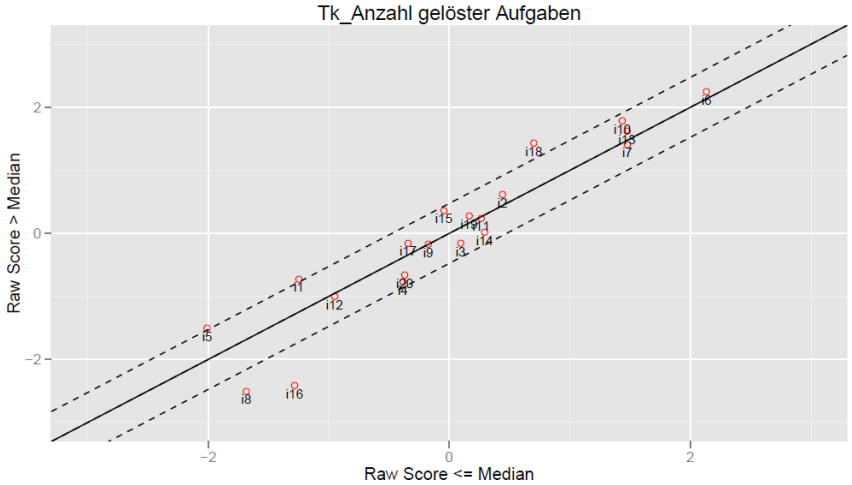


Abbildung 15. Brasilien

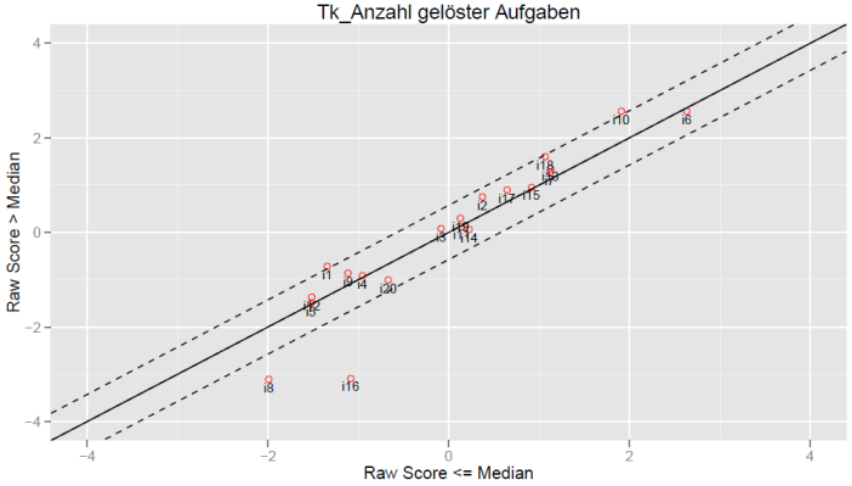


Abbildung 16. Kanada

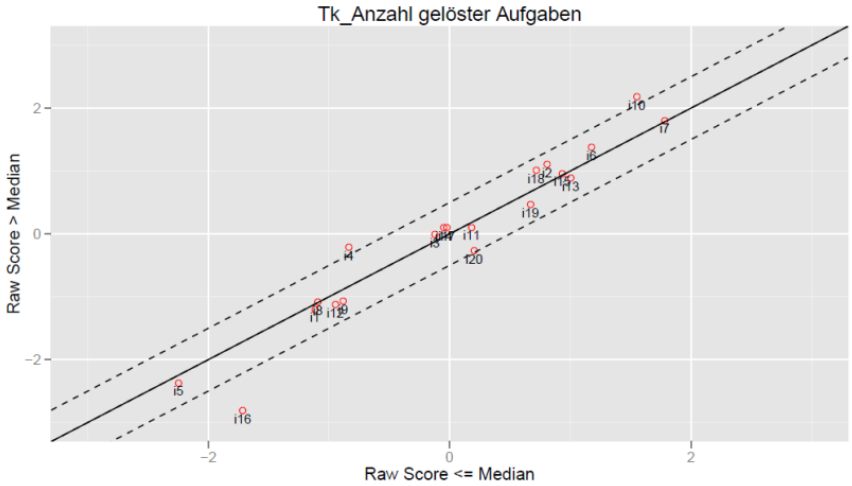


Abbildung17. Tschechische Republik

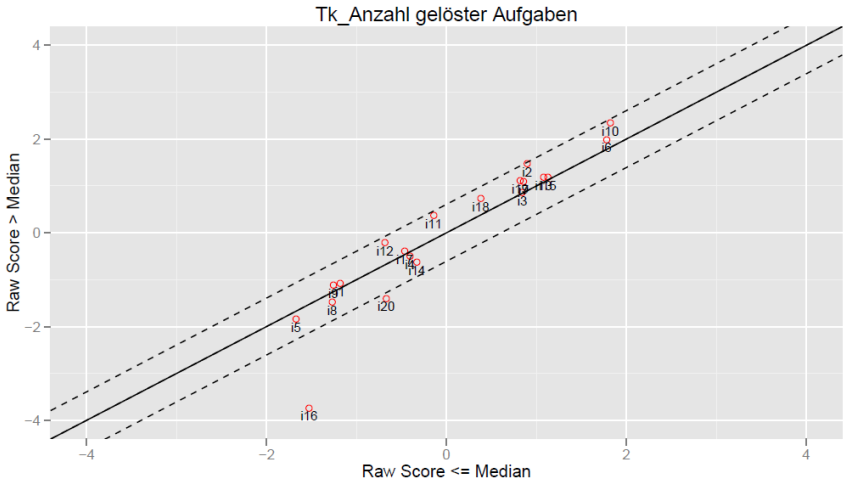


Abbildung18. Dänemark

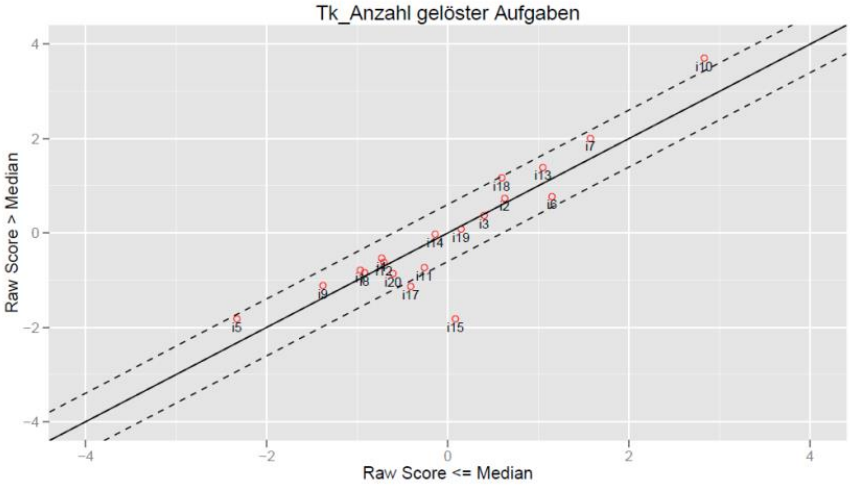


Abbildung 19. Deutschland

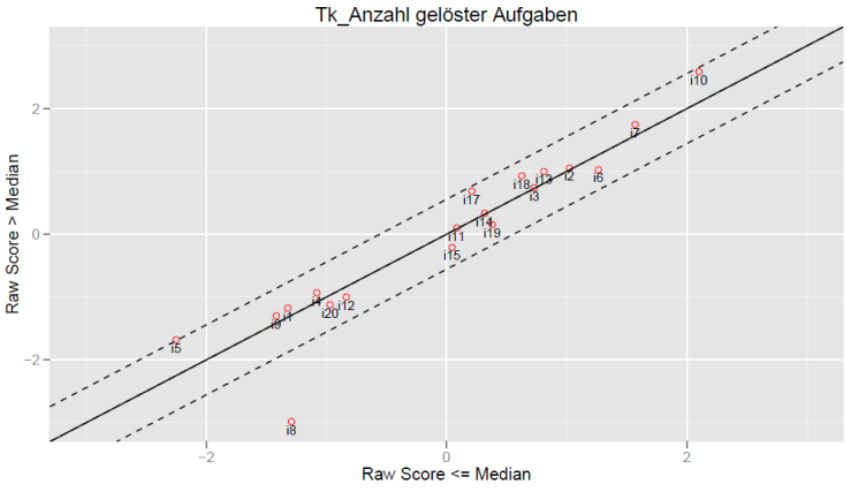


Abbildung 20. Finnland

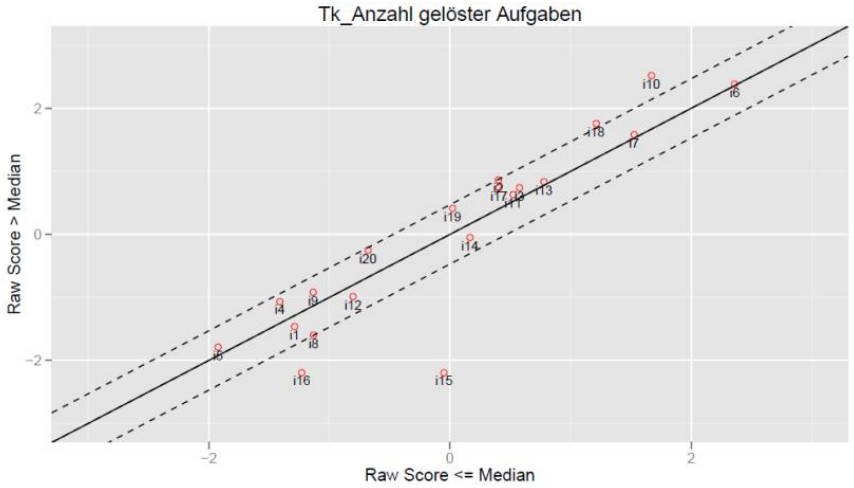


*Abbildung 2*1. Frankreich

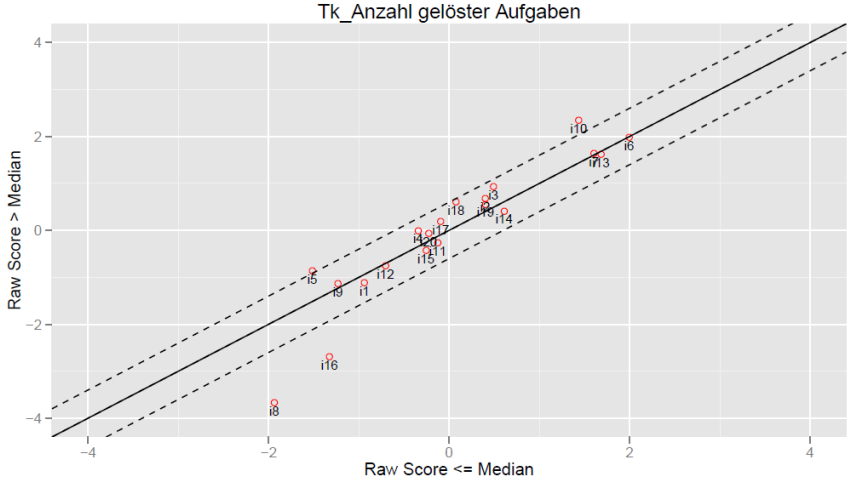


Abbildung 22. Griechenland

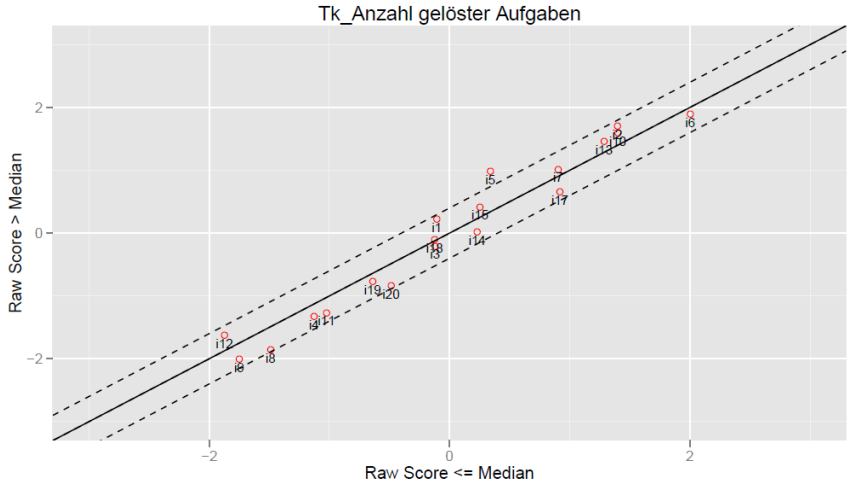


Abbildung23. Hong Kong - China

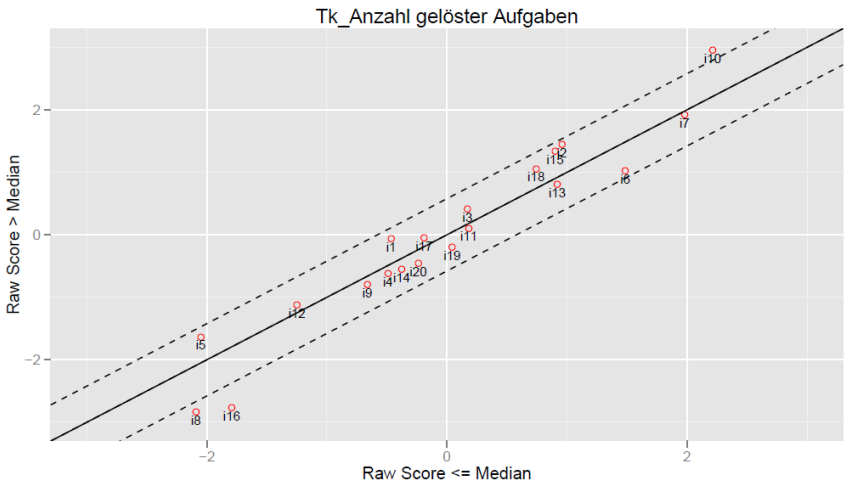


Abbildung 24. Ungarn

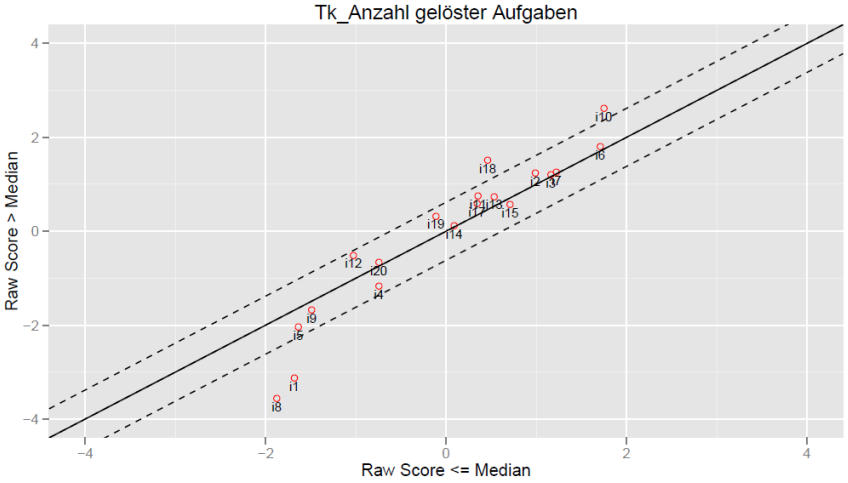


Abbildung 25. Island

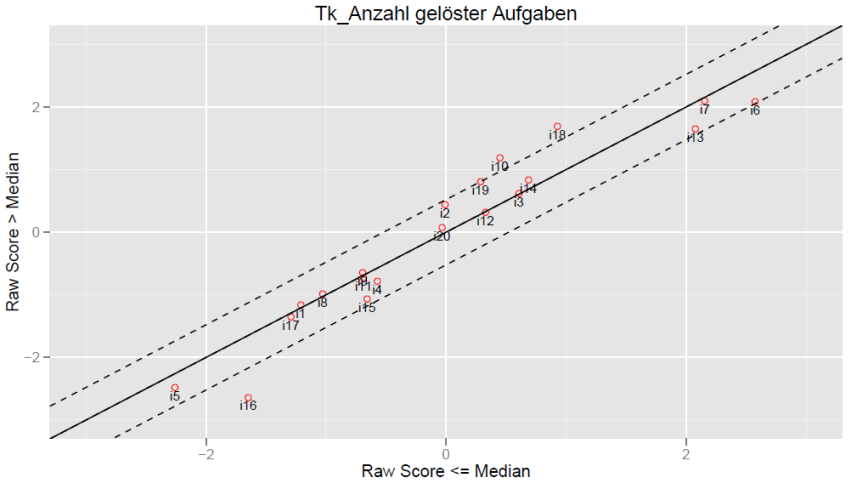


Abbildung 26. Indonesien

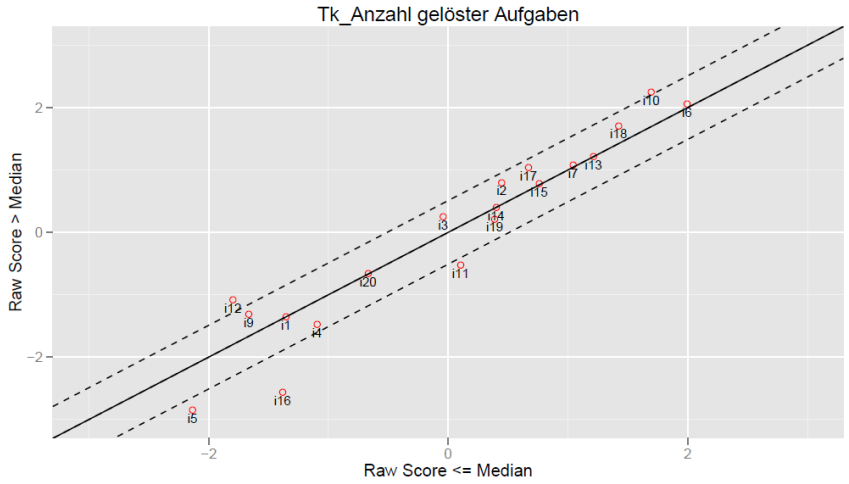


Abbildung 27. Irland

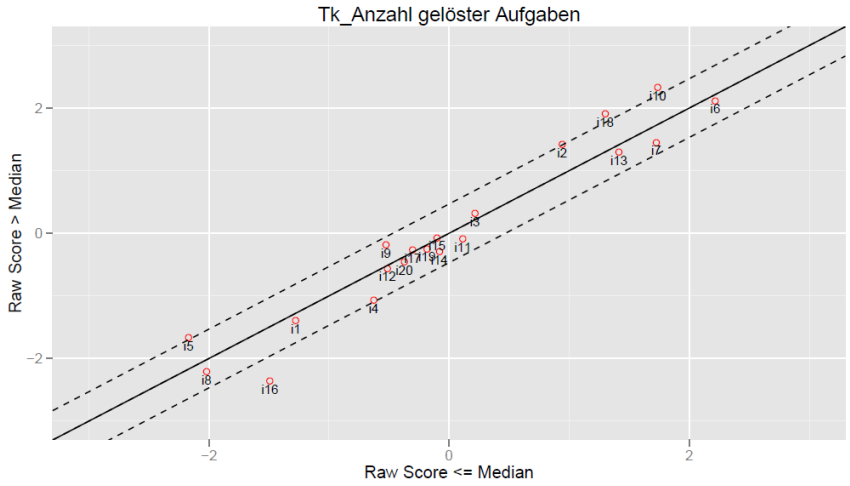
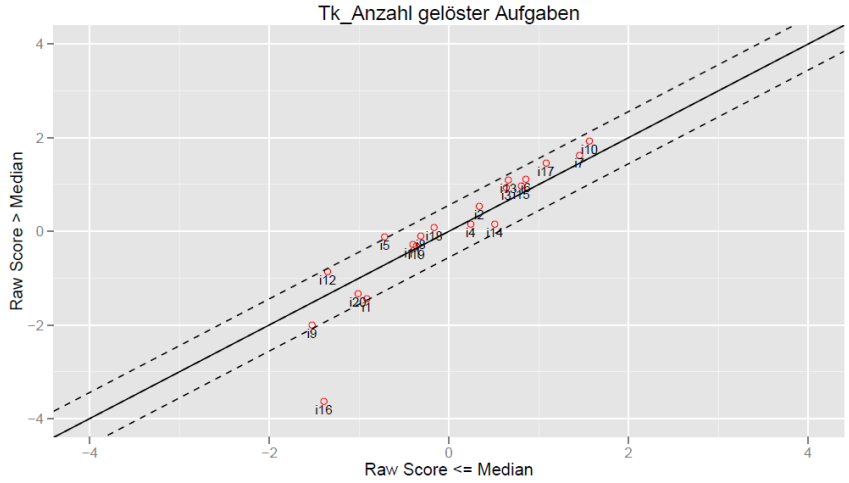


Abbildung 28. Italien



Abbilung 29. Japan

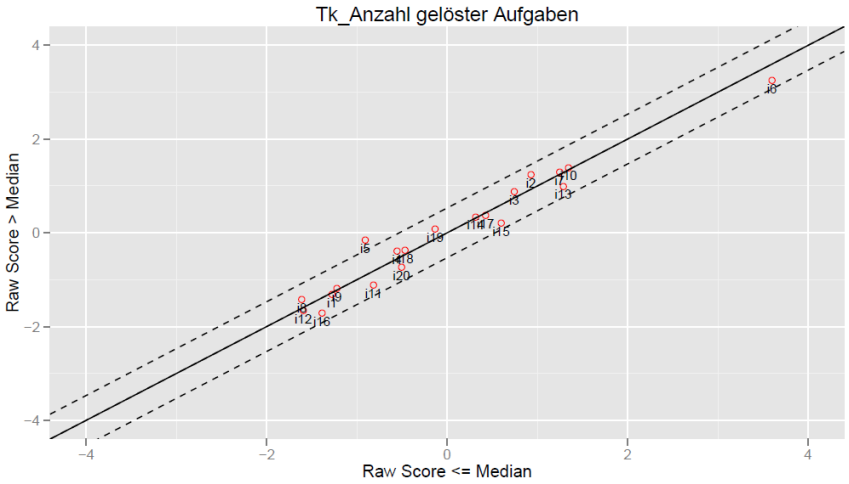


Abbildung 30. Korea

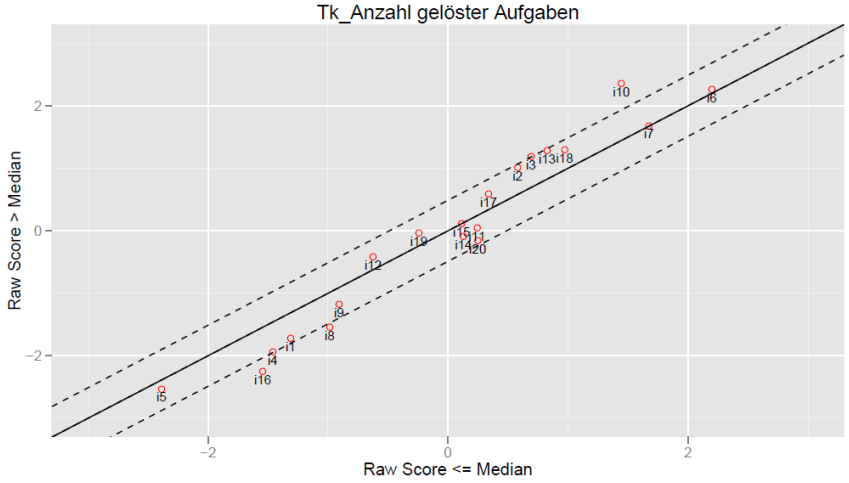


Abbildung 31. Lettland

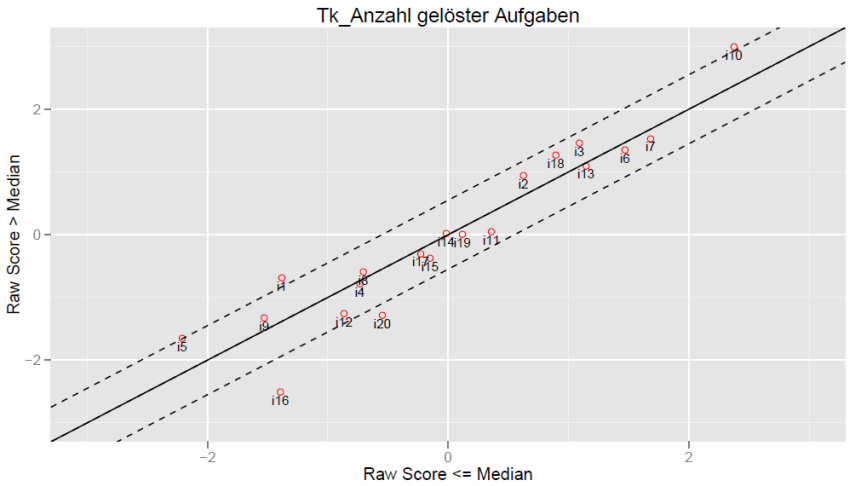


Abbildung 32. Luxembourg

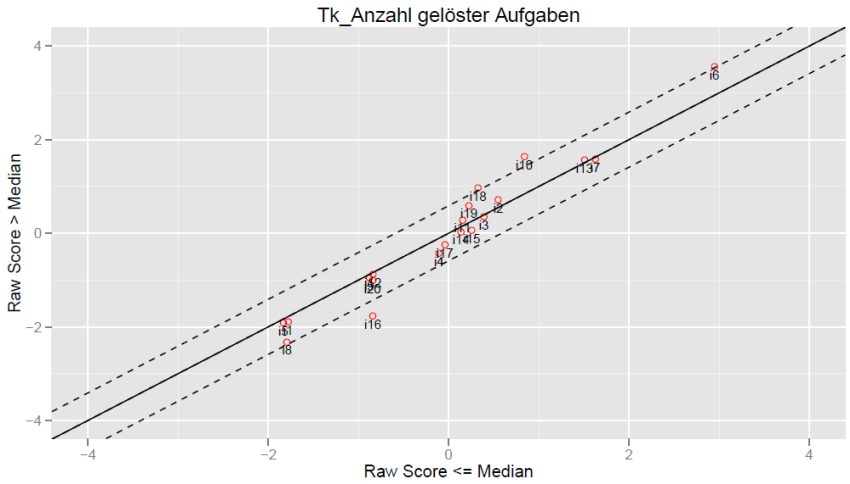


Abbildung 33. Mexiko

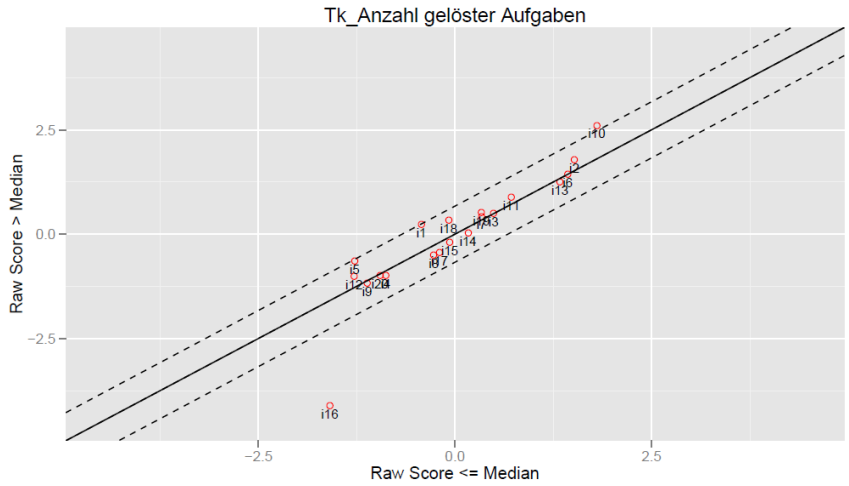


Abbildung 34. Niederlande

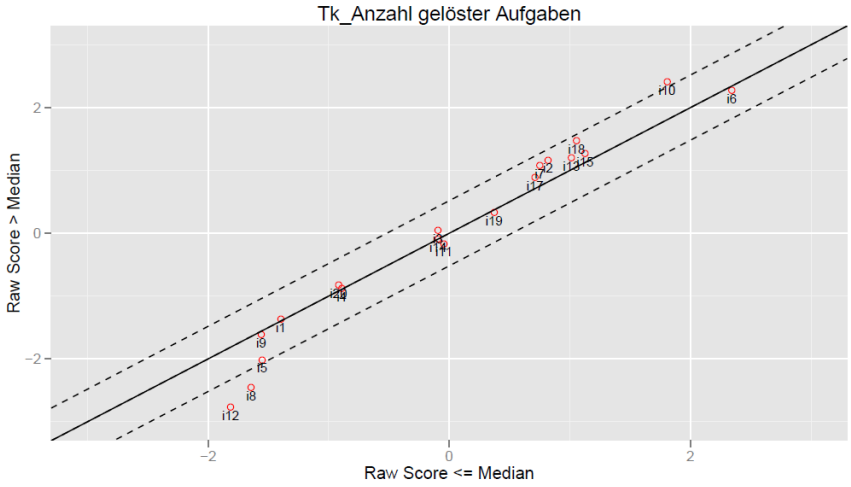


Abbildung 35. Neuseeland

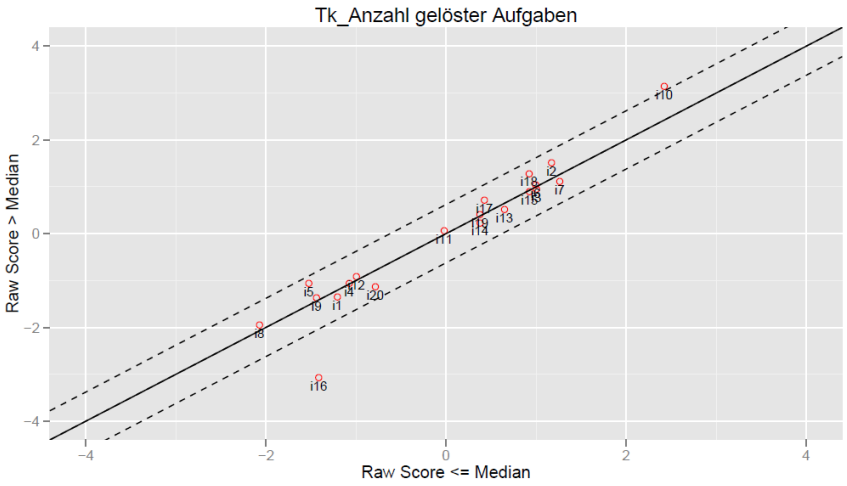


Abbildung 36. Norwegen

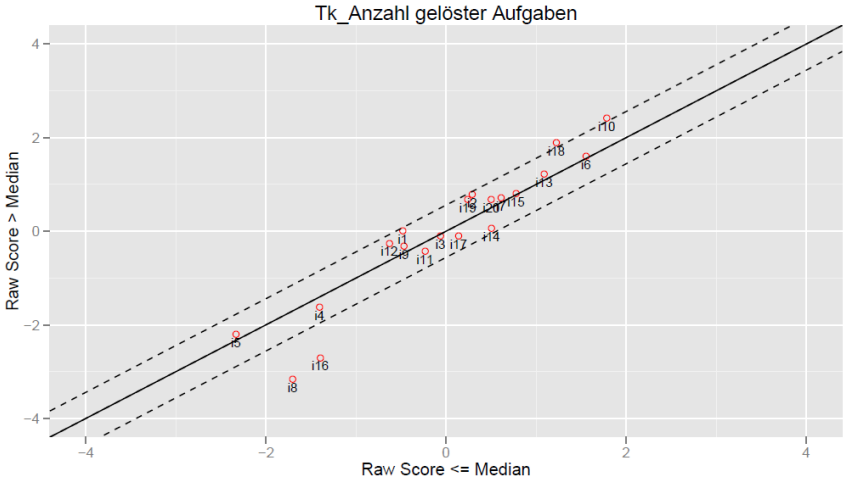


Abbildung 37. Polen

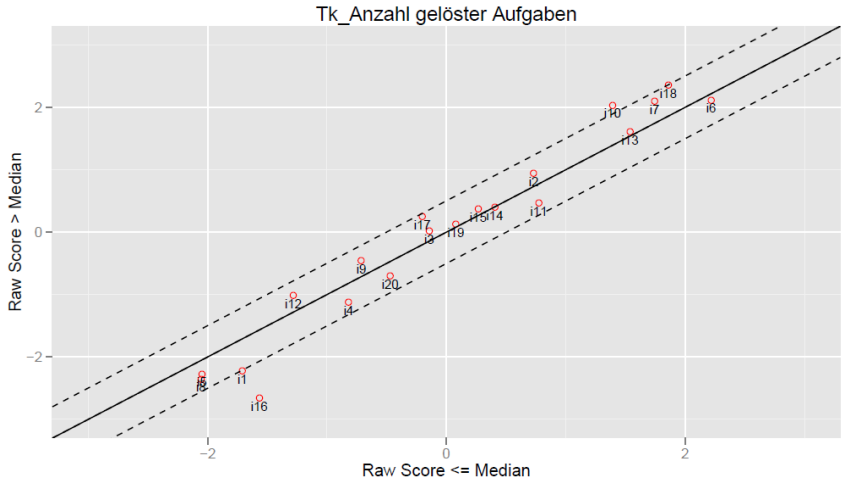


Abbildung 38. Portugal

Abbildung 39. Russland

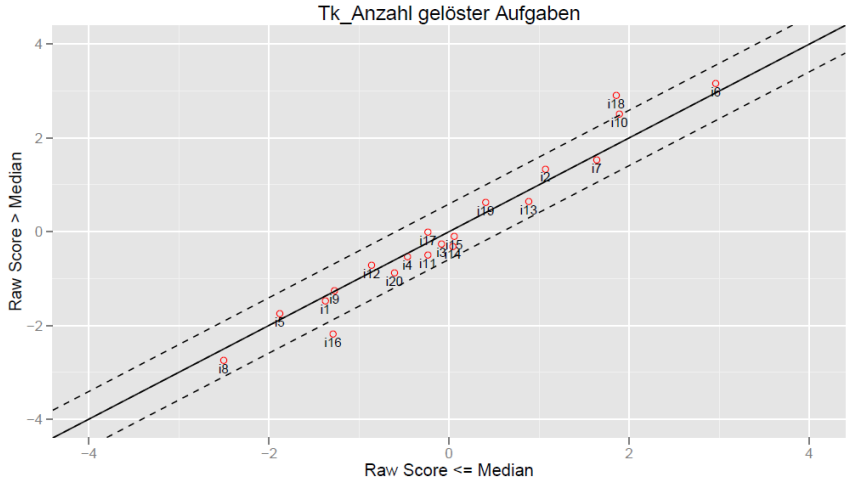


Abbildung 40. Spanien

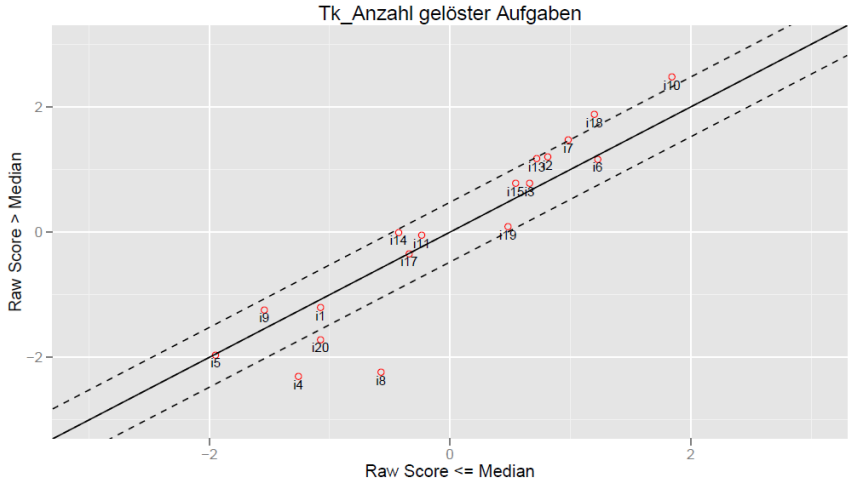


Abbildung 41. Schweden

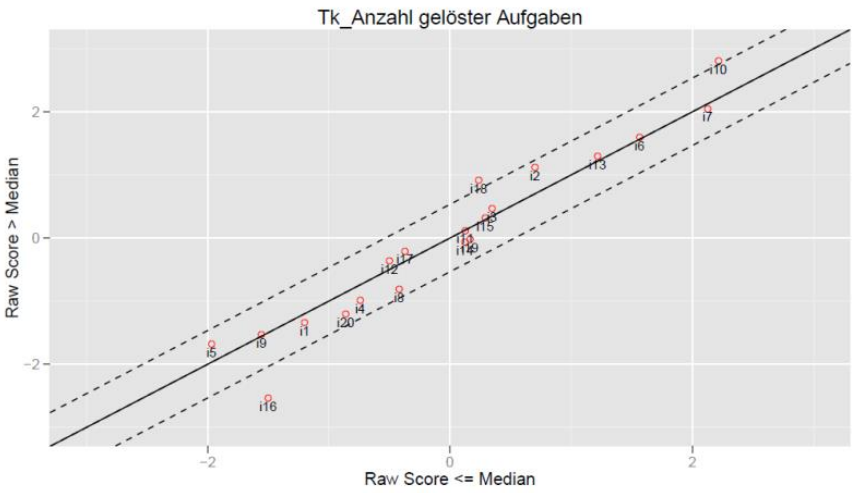


Abbildung 42. Schweiz

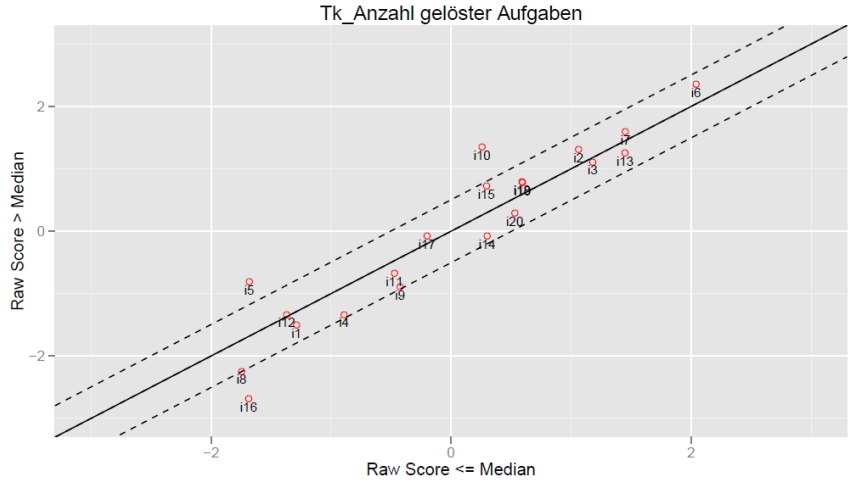


Abbildung 43. Thailand

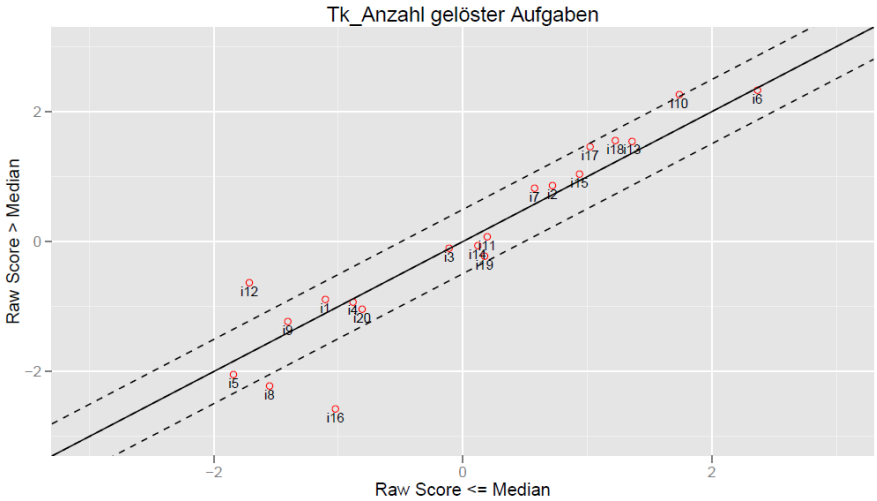


Abbildung 44. Großbritannien

Im Folgenden finden sich die Abbildungen der graphischen Modellkontrollen der einzelnen Länder in Bezug auf das Teilungskriterium Geschlecht.

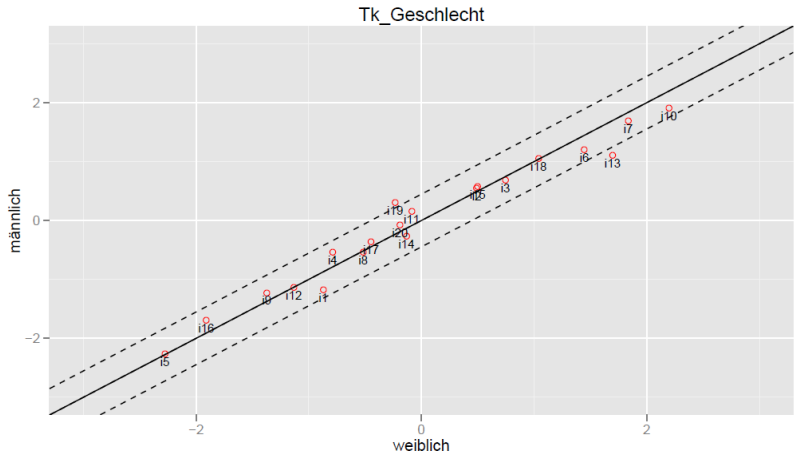


Abbildung45. Österreich

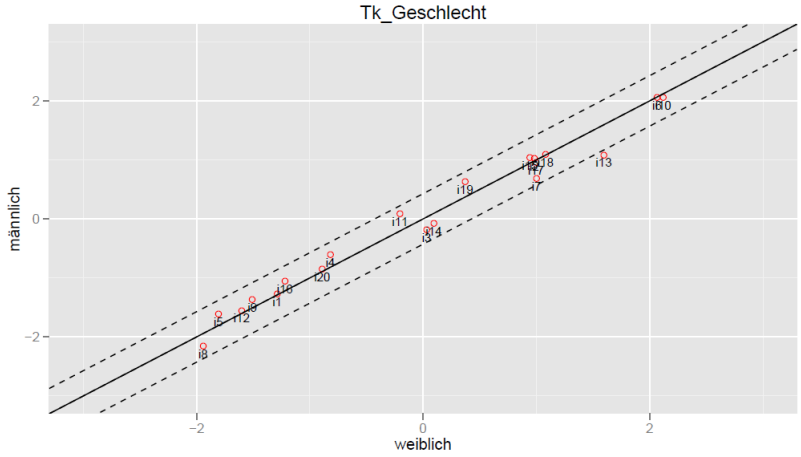


Abbildung 46. Australien

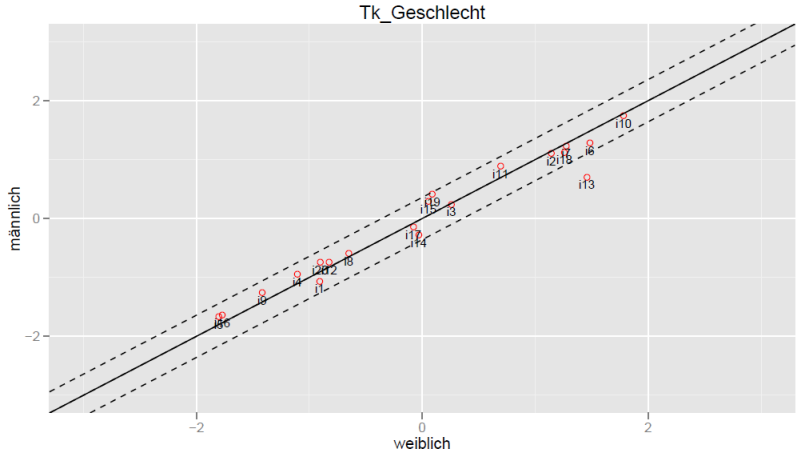


Abbildung 47. Belgien

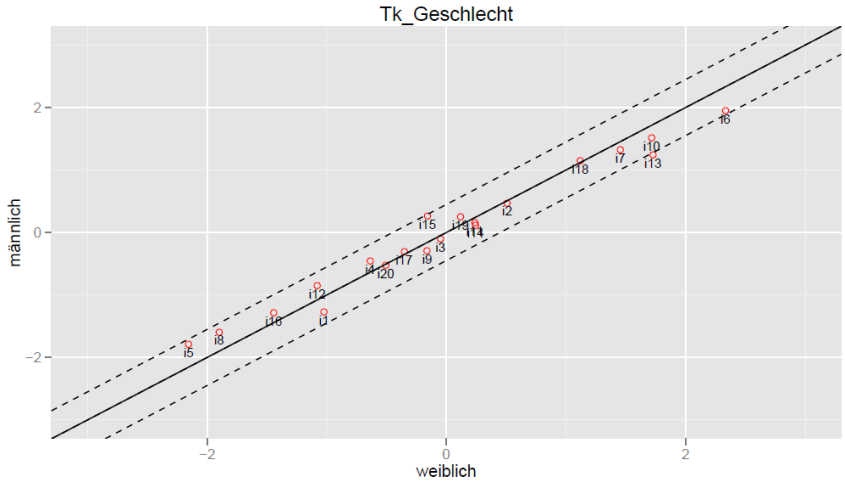


Abbildung 48. Brasilien

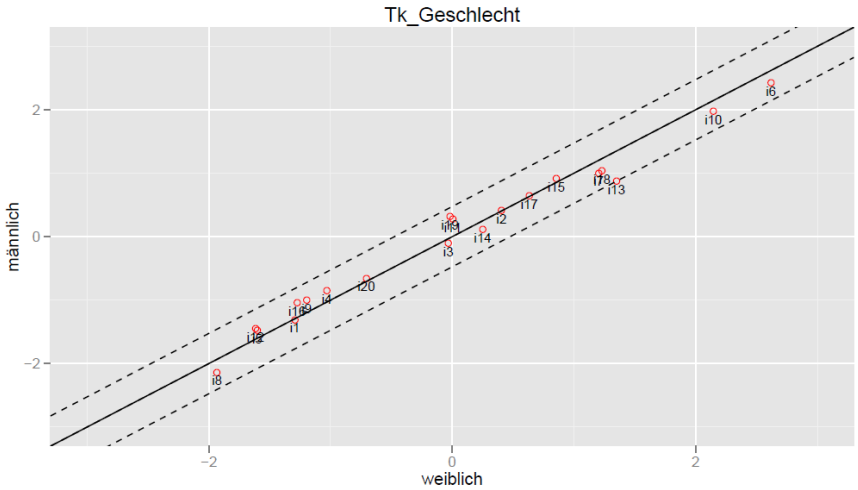


Abbildung 49. Kanada

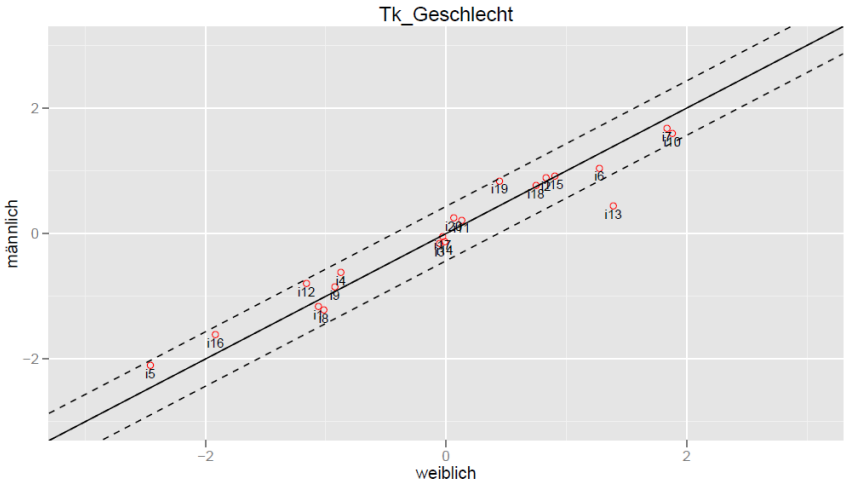


Abbildung 50. Tschechische Republik

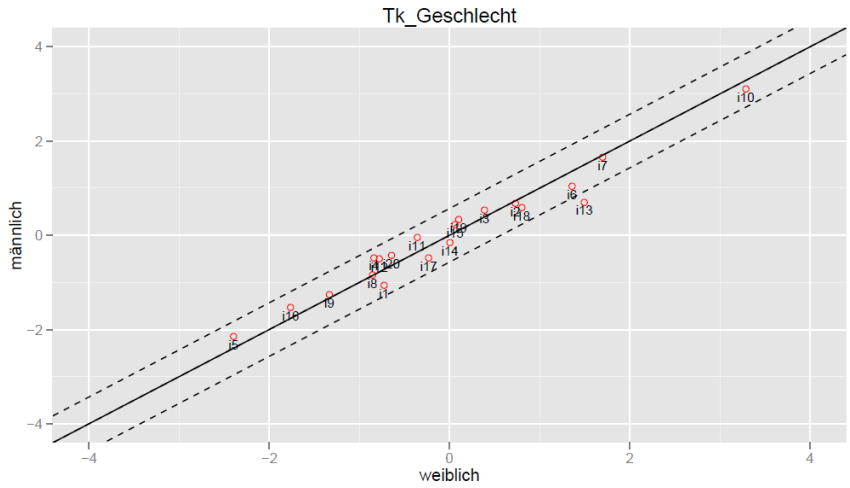


Abbildung 51. Deutschland

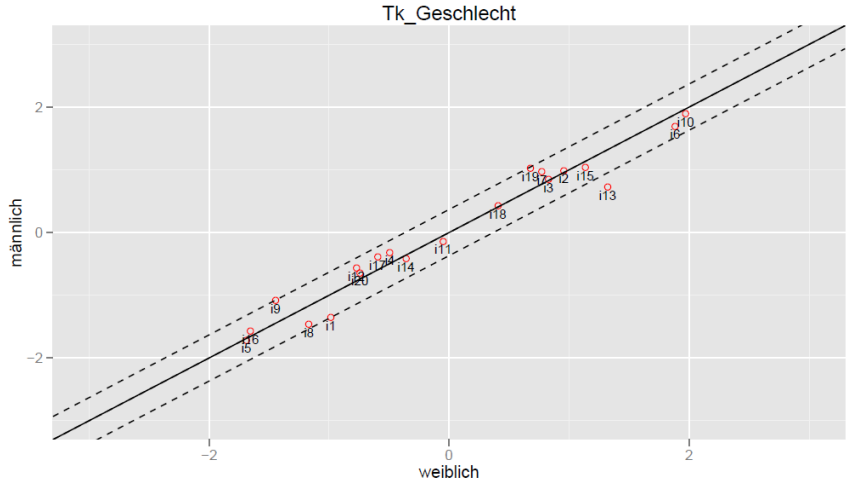


Abbildung 52. Dänemark

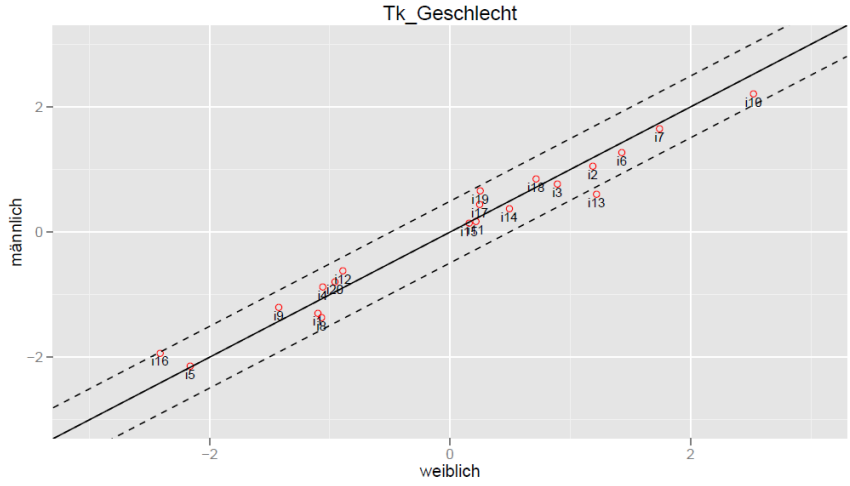


Abbildung 53. Finland

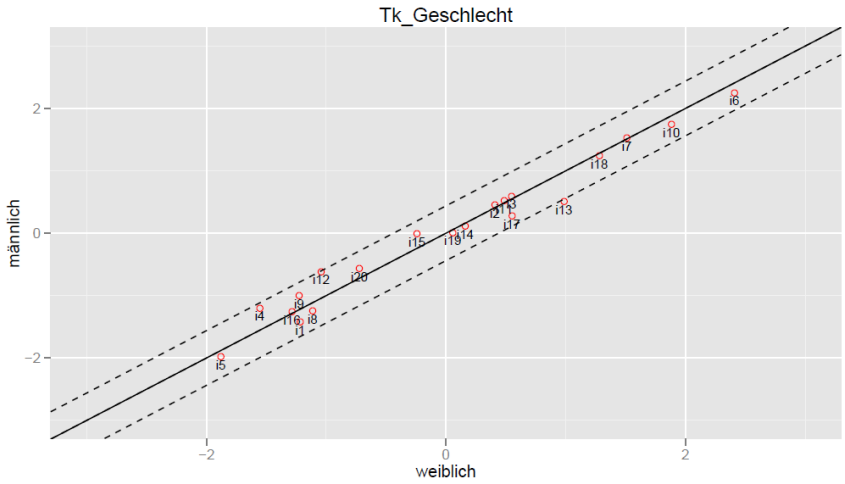


Abbildung 54. Frankreich

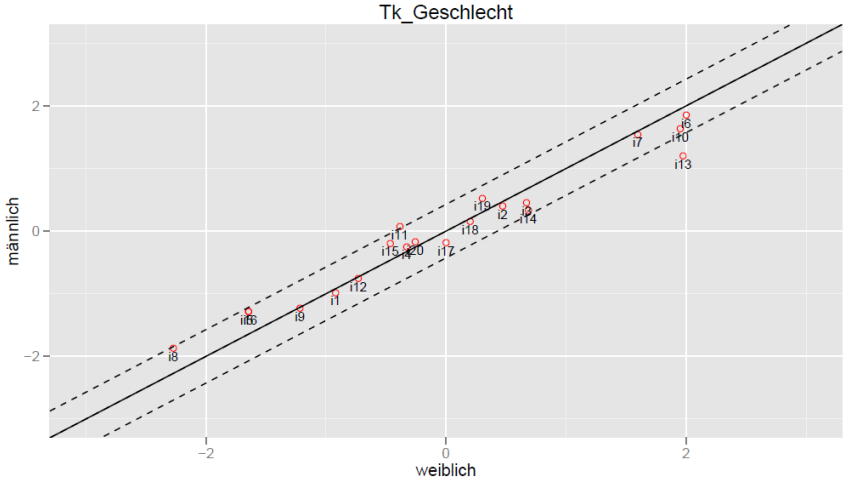


Abbildung 55. Griechenland

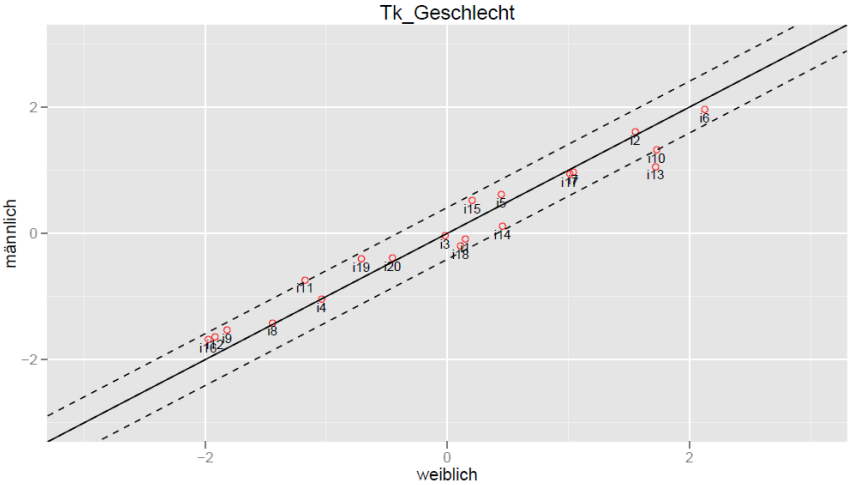


Abbildung 56. Hong Kong - China

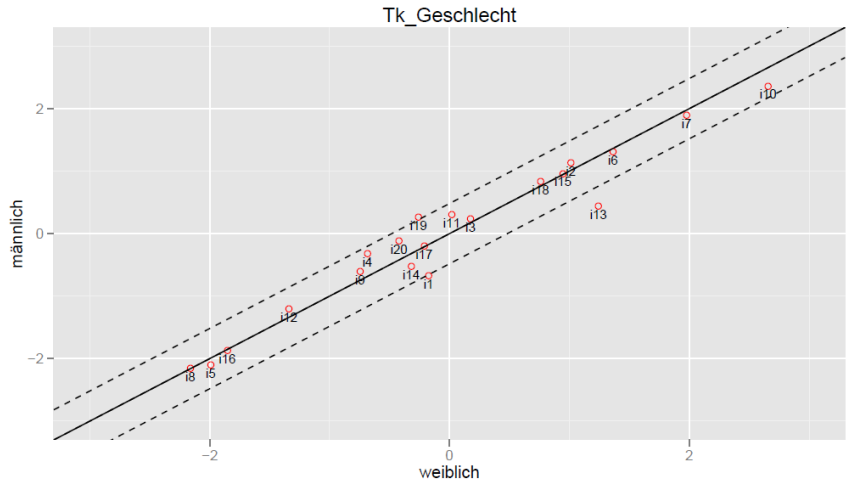


Abbildung 57. Ungarn

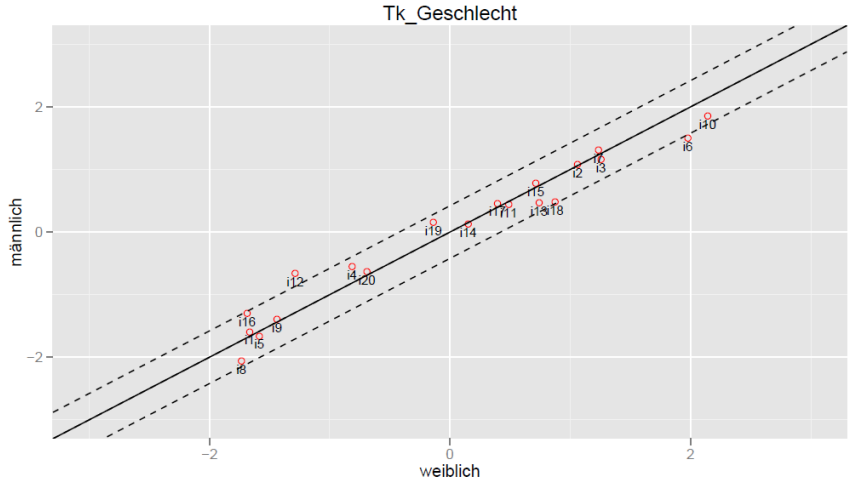


Abbildung 58. Island

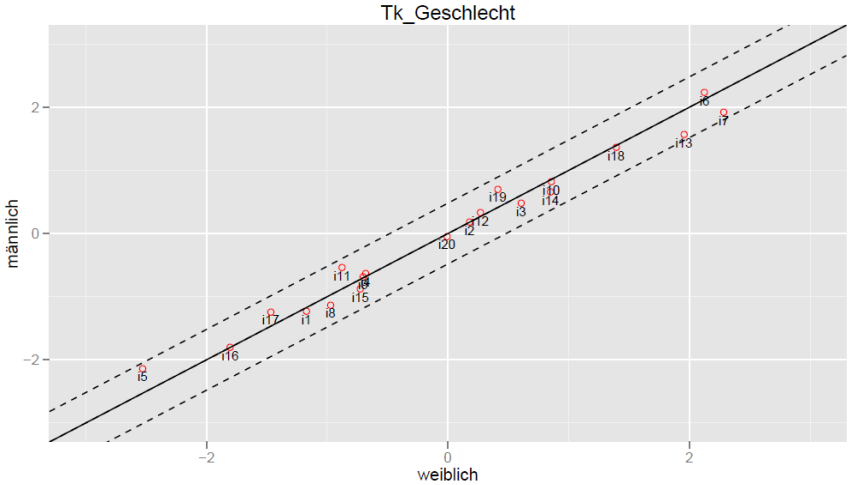


Abbildung 59. Indonesien

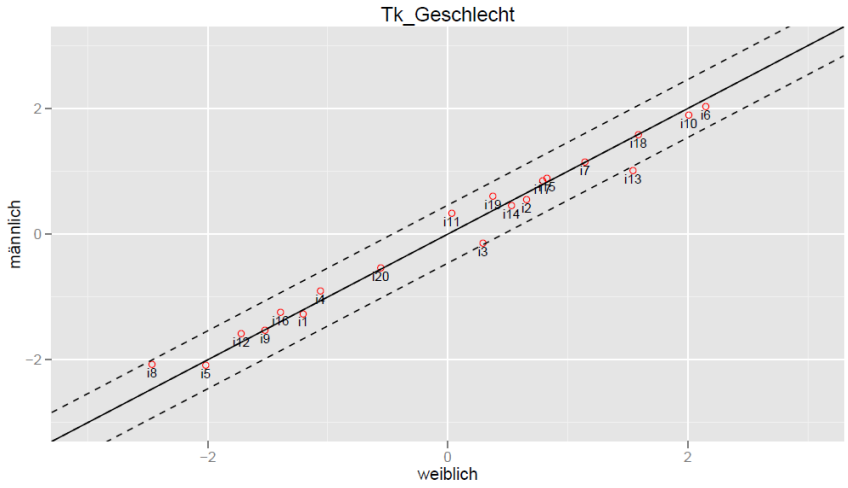


Abbildung 60. Irland

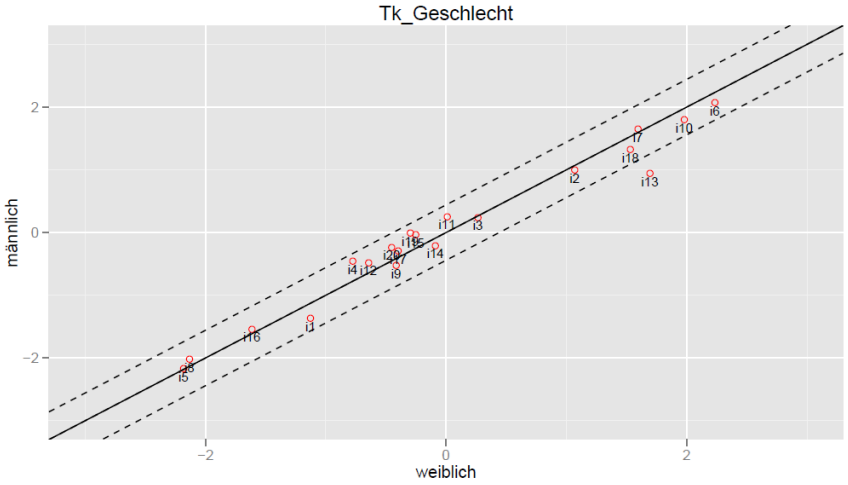


Abbildung 61. Italien

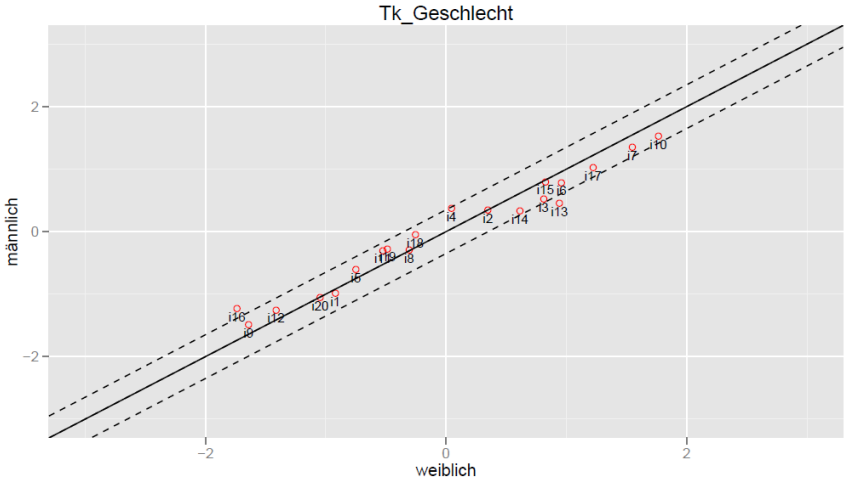


Abbildung 62. Japan

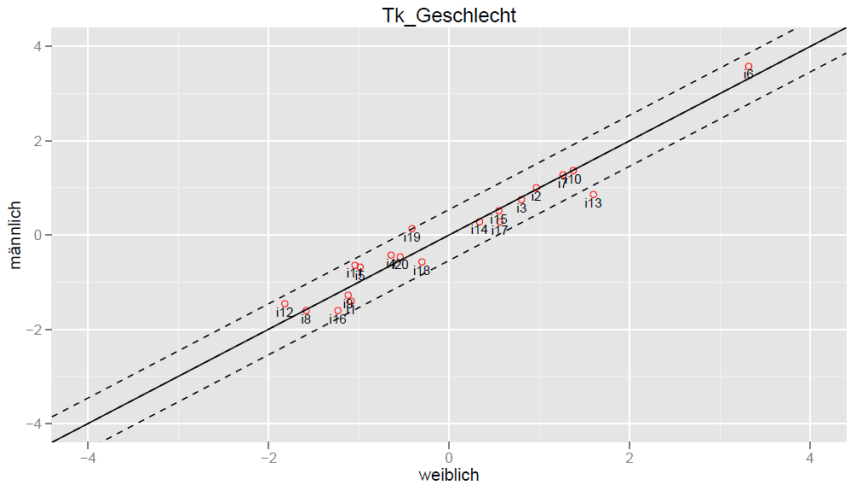


Abbildung 63. Korea

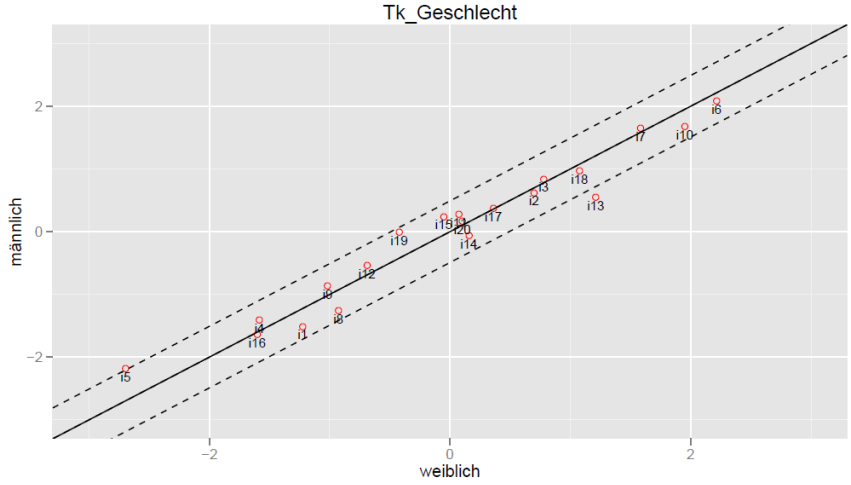


Abbildung 64. Lettland

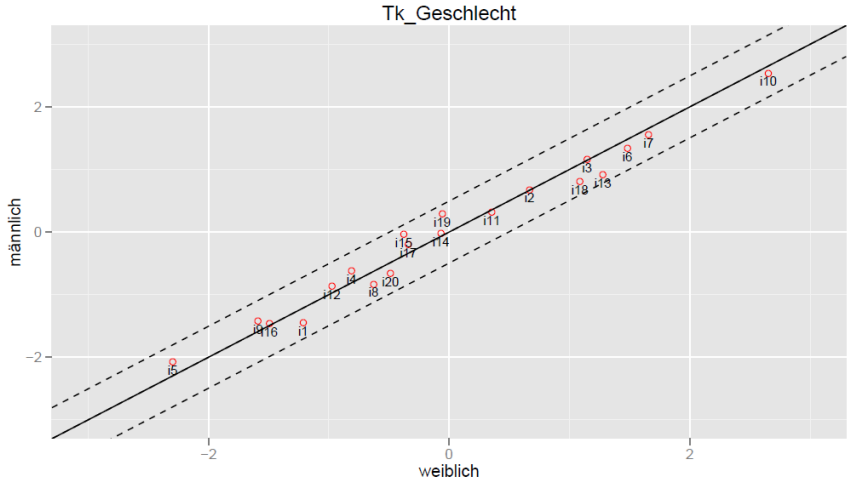


Abbildung 65. Luxembourg

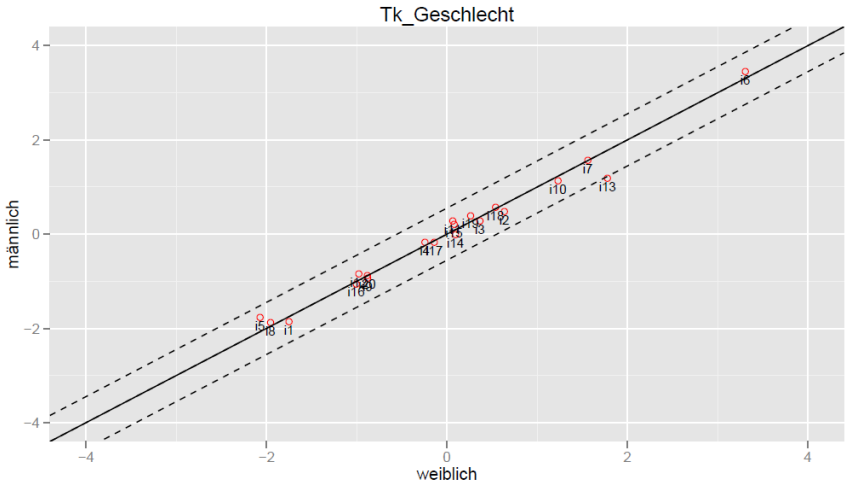


Abbildung 66. Mexiko

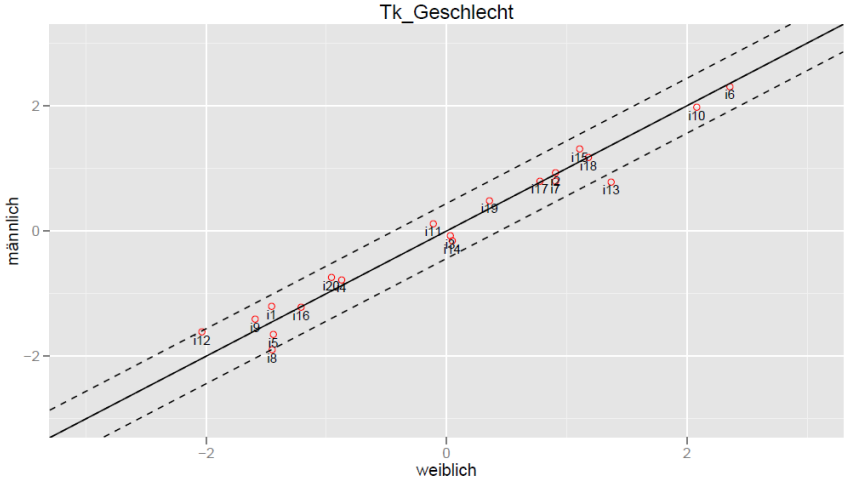


Abbildung 67. Neuseeland

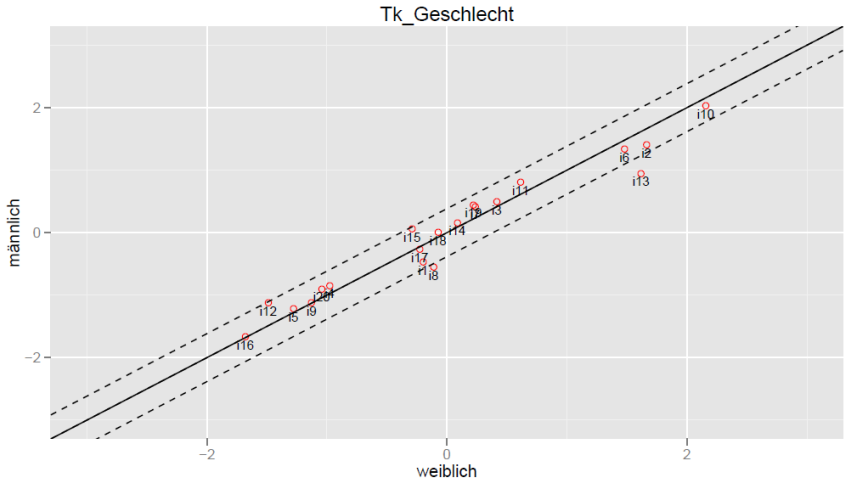


Abbildung 68. Niederlande

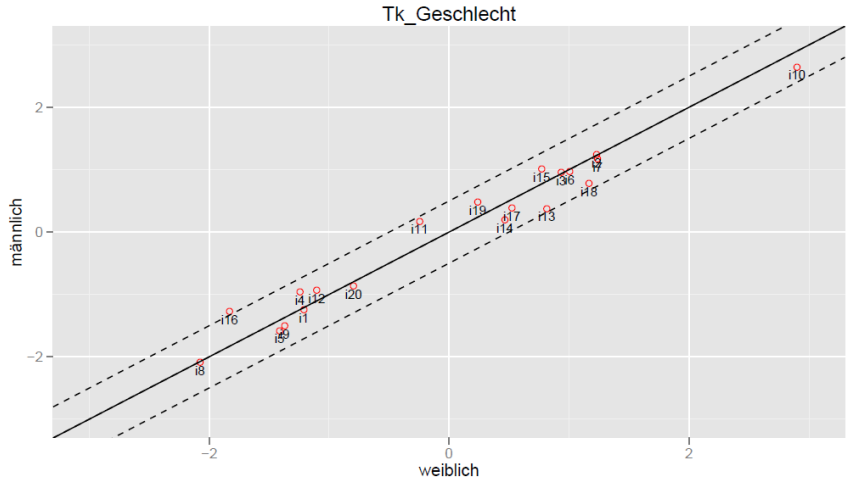


Abbildung 69. Norwegen

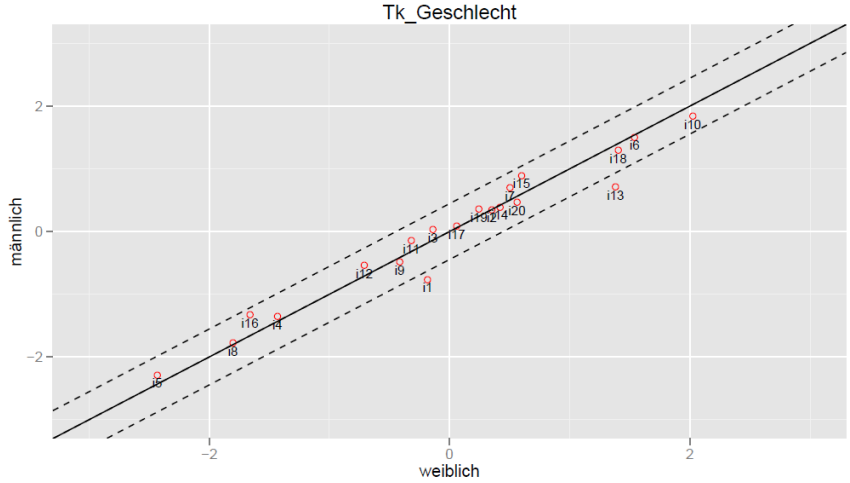


Abbildung 70. Polen

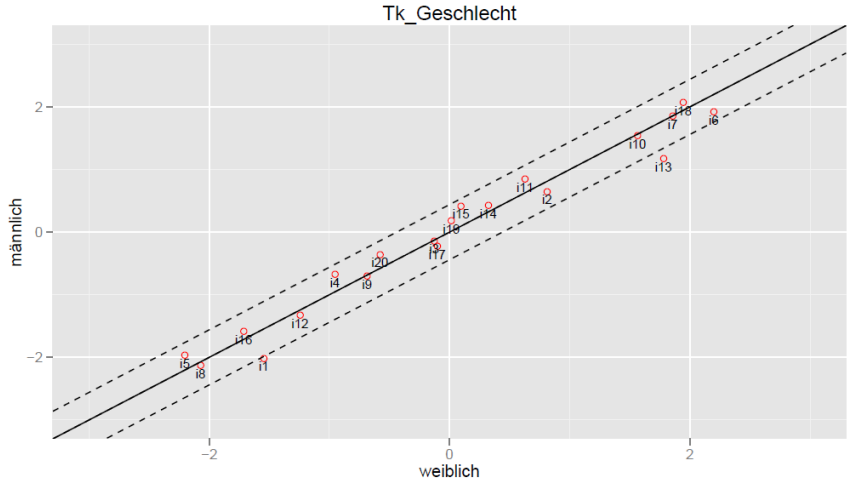


Abbildung 71. Portugal

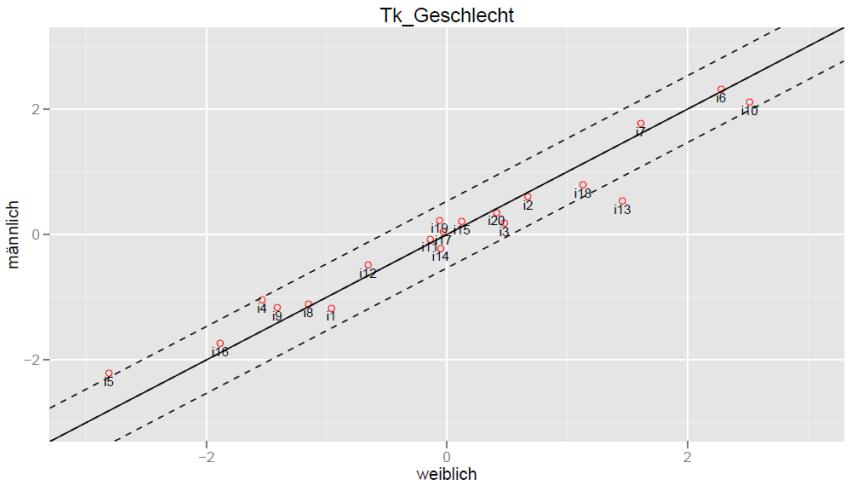


Abbildung 72. Russland

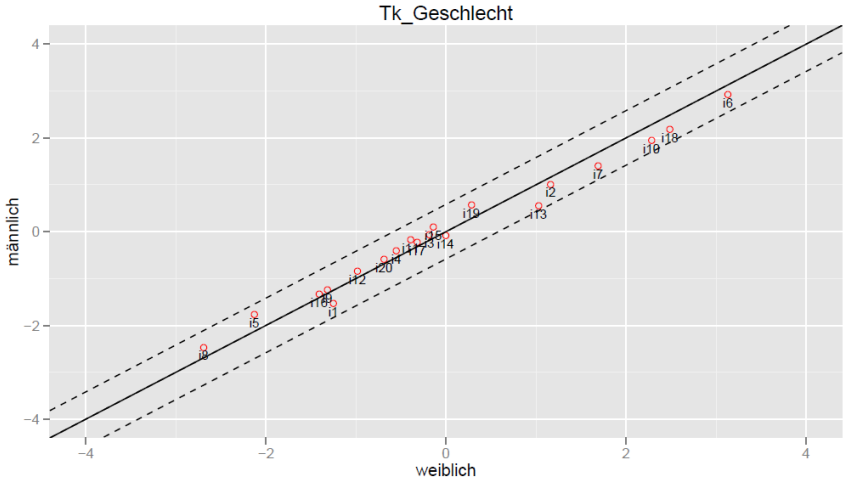


Abbildung 73. Spanien

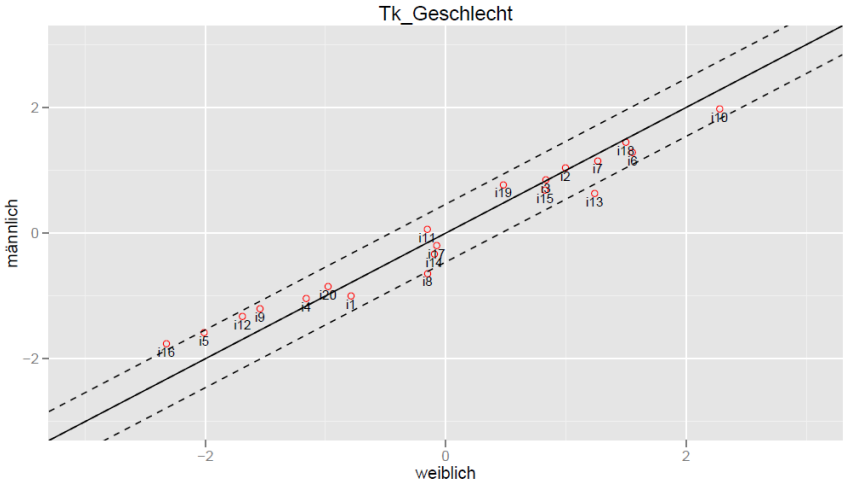


Abbildung 74. Schweden

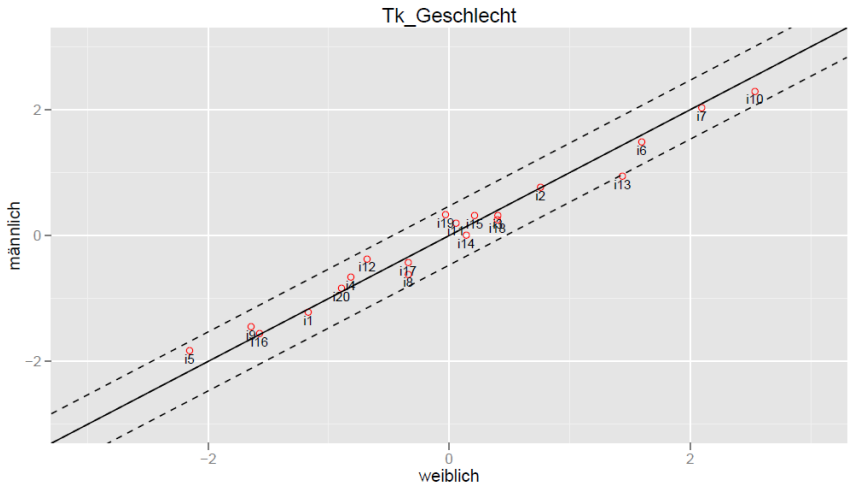


Abbildung 75. Schweiz

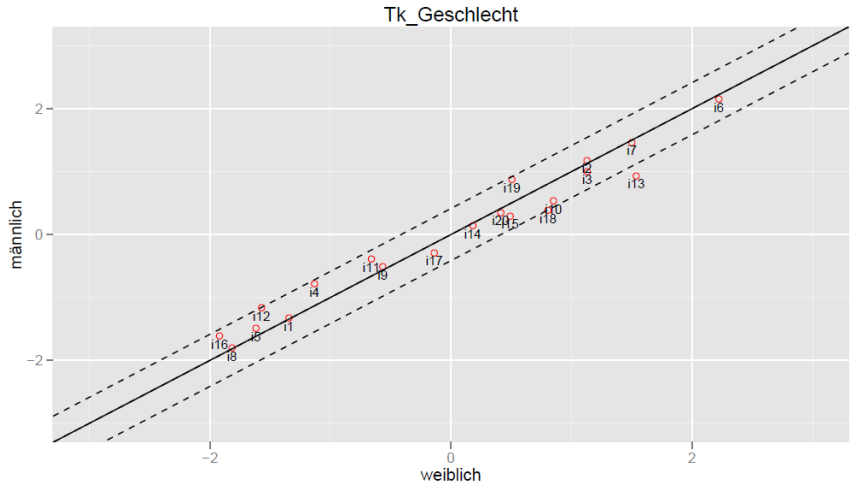


Abbildung 76. Thailand

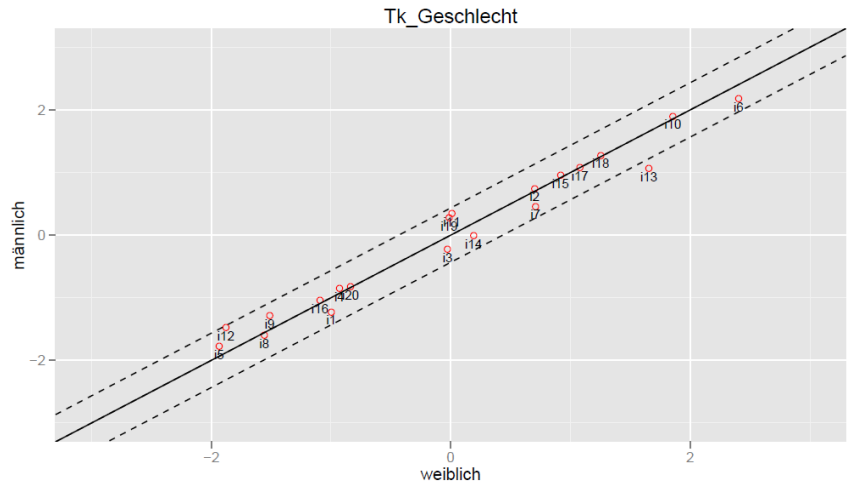


Abbildung 77. Großbritannien

9.2. Abstract

Deutsch

Diese Arbeit setzte sich zum Ziel die Items der PISA Studie (Programme for International Student Assessment) bezüglich ihrer Rasch-Modell-Konformität zu überprüfen. Analysiert wurden 20 Linking Items der PISA Studie – also jene Items, die in jedem Jahr vorgegeben wurden – aus den Jahren 2000, 2003, 2006 und 2009 aus dem Kompetenzbereich Lesen von 34 Ländern, die zu allen Zeitpunkten an der Studie teilgenommen haben. Insgesamt wurde ein Datensatz von 34 Ländern, 20 Items und 315 072 SchülerInnen ausgewertet. Die Auswertung wurde mittels Andersen-Likelihood-Ratio-Test (LRT; Andersen, 1973) für jedes Jahr und für jedes Land mit dem Teilungskriterium *Anzahl gelöster Aufgaben* und *Geschlecht* durchgeführt. Da es wahrscheinlich war, bereits auf Grund der ungewöhnlich großen Stichprobe, signifikante Ergebnisse zu erhalten (z.B. Kubinger, 2005), wurden für diese Arbeit nicht signifikante, sondern praktisch relevante Abweichungen als Kriterium zur Überprüfung des Rasch-Modells herangezogen. Für alle signifikanten Ergebnisse des LRT wurden daher graphische Modellkontrollen durchgeführt (Fischer, 1974). Eine Abweichung von der 45°-Geraden in der graphischen Modellkontrolle wurde als praktisch relevant angesehen, wenn die Differenz der Itemparameterschätzungen aus zwei Teilstichproben mehr als ein Zehntel der Spannweite der Parameterschätzungen betrug (Goethals, 1994). Die Analysen ergaben, dass beinahe in jedem Land bzw. in jedem Jahr und in Bezug auf beide Teilungskriterien Items vorhanden sind, die in den Teilgruppen nicht dieselbe Schwierigkeit aufweisen. Somit kann nicht davon ausgegangen werden, dass die Items der PISA Studie im Kompetenzbereich Lesen Rasch-Modell konform sind.

Englisch

The aim of this study was to check the Rasch model conformity used in the Programme for International Student Assessment (PISA). We analyzed a sample of linking items (recurring items in 2000, 2003, 2006 and 2009) of the PISA reading survey. The sample consisted of 34 countries, 20 items and 315 072 students. The Rasch model conformity was checked by testing the item parameters uniformity for students with low scores and with high scores and for males and females in every country and every year with Andersen's (1973) conditional likelihood ratio test (CLR). To address the statistical phenomenon of significance without relevance due to big sample sizes, a graphical model check was applied (e.g. Kubinger, 2005). A deviation in the graphical model check was considered to be practically relevant when the difference of two item parameter estimations of a subsample exceed a tenth of the range of the parameters (Goethals, 1994). Results show a practical relevant difference in item difficulty in almost every subsample in every country and every year. These results do not support the conformity of the Rasch model in the PISA reading survey.

9.3. Lebenslauf

Angaben zur Person

Name: Zabransky Nina

Adresse: Vorgartenstr.145-157/6/29, 1020 Wien

Geburtsdatum & -ort: 29.10.1988; Burgschleinitz (NÖ)

Kontakt: nina.za@gmx.net

Ausbildung

seit Oktober 2008	Diplomstudium Psychologie an der Universität Wien Abschluss des ersten Abschnitts Feb. 2011
1999 - 2007	Neusprachliches Gymnasium Horn Abschluss: Matura mit ausgezeichnetem Erfolg

---

Arbeitserfahrung

01.10.2014 – 31.01.2015	Studienassistenz an der Universität Wien Arbeitsbereich Sportpsychologie
01.12.2013 – 28.02.2014	Projektmitarbeit an der Universität Wien Arbeitsbereich Sportpsychologie
19.11.2012 – 19.12.2012	Projektmitarbeit an der Universität Wien Arbeitsbereich Sportpsychologie
03.09.2012 – 23.11.2012	Praktikum an der Universität Wien Arbeitsbereich Sportpsychologie

---

Persönliche Fähigkeiten und Kompetenzen

Sprachkenntnisse: Deutsch. Englisch, Französisch, Latein

Computerkenntnisse: MS Office, SPSS, R