Technischer Bericht zum Check S2 2019

Stéphanie Berger, Laura A. Helbling, Nina König, Martin J. Tomasik, Urs Moser

Institut für Bildungsevaluation (IBE)
Assoziiertes Institut der Universität Zürich

Zürich, 22. November 2019

Inhaltsverzeichnis

1 Einleitung

Die Kantone Aargau, Basel-Stadt, Basel-Landschaft und Solothurn haben das Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich, mit der Entwicklung und Durchführung gemeinsamer Leistungstests in der 3. und 5. Klasse der Primarstufe (Check P3 und Check P5) sowie in der 2. und 3. Klasse der Sekundarstufe I (Check S2 und Check S3) beauftragt. Die sogenannten Checks prüfen fachliche Leistungen in Deutsch, Englisch, Französisch, Mathematik sowie Natur und Technik. Die Checks werden als externe standardisierte Standortbestimmung durchgeführt, mit dem Ziel, den Schülerinnen und Schülern eine unabhängige klassenübergreifende Beurteilung ihrer Kompetenzen zur Verfügung zu stellen. Die Ergebnisse werden in Bezug zum Lehrplan 21 (kriterienbezogene Norm) und im Vergleich zu allen Schülerinnen und Schülern des Kantons beziehungsweise des Bildungsraums Nordwestschweiz (Sozialnorm) auf Individual-, Klassen- und Schulebene zurückgemeldet. Zudem können Schülerinnen und Schüler ihren Lernfortschritt (individuelle Bezugsnorm) zwischen der 2. und 3. Klasse der Sekundarstufe I anhand der Checks S2 und S3 nachvollziehen. Die unabhängige kompetenzorientierte Leistungsbeschreibung dient in erster Linie dem gezielten Fördern und Lernen im Unterricht sowie der Unterrichts- und Schulentwicklung. Dieser technische Bericht bezieht sich auf den Check S2 2019.

2 Testdesign

Der Check S2 2019 umfasste insgesamt zehn Tests, wobei zwei verschiedene Testformate zu unterscheiden sind: Online-Tests und Papier-Tests. Tabelle 1 enthält eine Übersicht über die erfassten Kompetenzbereiche, die Testformate sowie die Anzahl Aufgaben für die Online-Tests pro Test oder Testteil.

2.1 Online-Tests

Die Online-Tests wurden am Computer durchgeführt und waren in der Regel als adaptive Multistage-Tests (Duanli, von Davier & Lewis, 2014) konzipiert. In Deutsch, Englisch und Französisch wurden je zwei Online-Tests für die folgenden Kompetenzbereiche durchgeführt: *Deutsch Sprache im Fokus*, *Deutsch Lesen*, *Englisch Lesen*, *Englisch Hören*, *Französisch Lesen* und *Französisch Hören*. In Natur und Technik wurden ein allgemeiner Test sowie je zwei spezifische Tests durchgeführt. In Mathematik wurde ein Online-Test durchgeführt, in dem die Kompetenzbereiche *Zahl und Variable*, *Form und Raum* (Geometrie) sowie *Grössen, Funktionen, Daten und Zufall* gleichzeitig erfasst wurden.

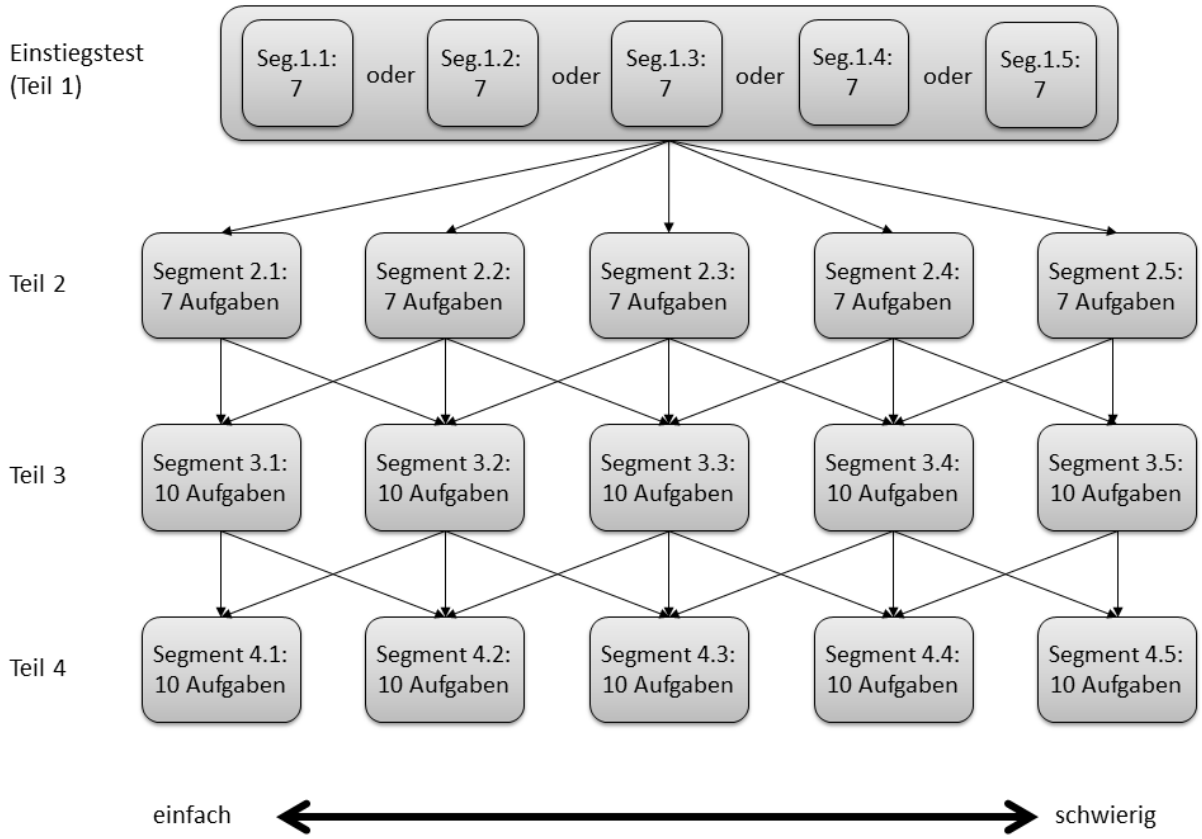
Die Online-Aufgaben lagen in unterschiedlichen Formaten wie Multiple-Choice, Lückentext, Aufzählung, Drag und Drop (z.B. Zuordnung von verschiedenen Begriffen zu Bildern) oder Hotspot (z.B. mit der Maus auf die richtige Stelle in einem Bild klicken) vor. Grundlegend handelte es sich bei allen gewählten Formaten um Aufgaben, die sich in richtig oder falsch dichotomisieren lassen. Abbildung 1 zeigt exemplarisch das Testdesign für den Online-Test *Deutsch Sprache im Fokus*. Dieses Testdesign gilt (ausser für Natur und Technik) für alle Skalen, die online getestet wurden. Ledig-

Tabelle 1: Kompetenzbereiche, Testformate und Anzahl Aufgaben pro Test

	Format	Aufgaben pro Testteil				
Fach		Teil 1	Teil 2	Teil 3	Teil 4	Total
Mathematik	Online	9	9	15	15	48
Deutsch						
Lesen	Online	10	10	10	10	40
Sprache im Fokus	Online	7	7	10	10	34
Schreiben	Papier	–	–	–	–	–
Französisch						
Lesen	Online	7	7	8	8	30
Hören	Online	7	7	8	8	30
Englisch						
Lesen	Online	8	8	8	8	32
Hören	Online	7	7	8	8	30
Schreiben	Papier	–	–	–	–	–
Natur und Technik[a]						
Allgemeine Naturwissenschaften	Online	10	8	8	–	26
Physik, Chemie oder Biologie	Online	5	5	–	–	10

[a] Der Test in Natur und Technik besteht aus einem allgemeinen Drei-Stage-Test mit 26 Aufgaben sowie zwei weiteren Zwei-Stage-Tests mit je 10 Aufgaben, welche aus den Fachbereichen Physik, Chemie oder Biologie ausgewählt werden können.

Abbildung 1: Multistage-Design des Online-Tests *Deutsch Sprache im Fokus* des Check S2



lich die Aufgabenzahl pro Testteil sowie die Anzahl an Einstiegssegmenten können zwischen diesen Online-Tests variieren.

Jeder Multistage-Test umfasste vier Testteile (Teil 1–4, vgl. Abbildung 1) und begann mit dem ersten Testteil (Einstiegstest), dessen Aufgaben einen eher einfacheren Schwierigkeitsgrad aufwiesen. Dieser erste Testteil bestand aus fünf äquivalenten Testsegmenten, wobei die Schülerinnen und Schüler zufällig einem davon zugewiesen wurden. Für die Testteile 2 bis 4 standen jeweils fünf Segmente unterschiedlicher Schwierigkeit zur Verfügung. Die Schülerinnen und Schüler wurden auf Basis ihrer Leistungen im vorangehenden Testteil adaptiv zu jenem Segment des nachfolgenden Testteils weitergeleitet, dessen Schwierigkeit am besten mit ihren Fähigkeiten übereinstimmte. Diese Abstimmung zwischen der Schwierigkeit der Testsegmente und der Fähigkeit der Schülerinnen und Schüler erlaubt es, die Fähigkeiten der Schülerinnen und Schüler effizient zu messen und den Messfehler zu minimieren (Duanli et al., 2014).[1]

[1]Im Anhang wird die Testinformation pro Hauptpfad (1–5) grafisch dargestellt. Hauptpfade beschreiben hier das vertikale Durchlaufen des Tests, wobei Schülerinnen und Schüler Testteil 2 bis 4 auf derselben Schwierigkeitsstufe lösen. Hauptpfad 1 beispielsweise entspricht einem Pfad, bei dem Schülerinnen und Schüler Testteil 2 bis 4 auf der leichtesten Schwierigkeitsstufe lösen. Hauptpfad 5 entspricht demgegenüber einem Pfad, bei dem die Testteile 2 bis 4 auf der höchsten Schwierigkeitsstufe gelöst wurden.

Die Aufgaben eines Testteils mussten innerhalb einer vorgegebenen Zeit gelöst werden. Je nach Testteil und Skala variierte diese Zeit zwischen 5 und 25 Minuten. Die Schülerinnen und Schüler hatten innerhalb eines Testteils und vor Ablauf der Zeit die Möglichkeit, Aufgaben auszulassen, zu Aufgaben zurückzukehren oder die Antwort zu korrigieren. Nach Ablauf der Zeit konnten die Aufgaben des Testteils nicht mehr weiterbearbeitet werden.

2.2 Online-Test in Natur und Technik

Der Check in Natur und Technik bestand aus drei Online-Tests. Der erste Online-Test war als adaptiver Drei-Stage-Test konzipiert und enthielt Aufgaben zum Thema «Allgemeine Naturwissenschaften». Der allgemeine Teil enthielt zwei äquivalente Einstiegssegmente, welche den Schülerinnen und Schülern zufällig zugeordnet wurden. Abhängig von den richtig gelösten Aufgaben standen den Schülerinnen und Schülern für die Testteile 2 und 3 jeweils drei unterschiedliche Schwierigkeitsstufen zur Verfügung («einfach», «mittel» und «schwierig»). Für den zweiten und dritten Online-Test standen je zwei Themen aus den Fachbereichen Biologie, Chemie und Physik zur Auswahl. Die Lehrperson entschied, welche zwei Themen in ihrer Klasse getestet wurden. Die Tests innerhalb der ausgewählten naturwissenschaftlichen Themen waren als adaptive Zwei-Stage-Tests konzipiert. Die Schülerinnen und Schüler erhielten zuerst fünf Aufgaben mittlerer Schwierigkeit und abhängig davon, wie gut sie diese lösten, wurden ihnen anschliessend fünf eher einfache oder fünf eher schwierige Aufgaben zugewiesen.

2.3 Papier-Tests

Für die Erfassung der Schreibkompetenzen wurden Papier-Tests eingesetzt. Die Schreibkompetenzen werden in Deutsch sowie alternierend in Englisch oder Französisch erfasst. Beim Papier-Test *Deutsch Schreiben* können die Schülerinnen und Schüler jeweils zwischen drei Themen wählen, die in ihrer Aufgabenstellung verschiedene Textsorten verlangen (u.a. Erzählung, Bildbeschreibung, Argumentation). Der Papier-Test zur Erfassung der Schreibkompetenzen in den Fremdsprachen Englisch oder Französisch wird jeweils in einer einfachen und in einer schwierigen Version angeboten. Welche der zwei Versionen eingesetzt wird, entscheidet die Lehrperson. Beide Versionen umfassen zwei Aufgaben (Schreibaufträge), die beide von den Schülerinnen und Schülern bearbeitet werden müssen. Wie in Deutsch implizieren die Aufgabenstellungen jeweils unterschiedliche Textsorten (u.a. Brief, Mitteilung, Erzählung, Reportage). Im Check S2 2019 wurden Schreibkompetenzen in Deutsch und Englisch getestet.

3 Aufgabenentwicklung

Alle Testaufgaben wurden von Fachdidaktikern und Fachdidaktikerinnen der Pädagogischen Hochschule Nordwestschweiz in Zusammenarbeit mit Lehrpersonen aus den vier Kantonen des Bildungsraums Nordwestschweiz entwickelt. Der Entwicklungsprozess umfasste diverse Feedbackrunden zur Beurteilung der Aufgaben aus fachdidaktischer und testtheoretischer Perspektive.

4 Teilnahme

Tabelle 2 gibt einen Überblick über die Anzahl Schulen, Klassen sowie Schülerinnen und Schüler, die pro Kanton am Check S2 2019 teilgenommen haben. Sie beinhaltet zudem die Anzahl zurückgemeldeter Online- und Papier-Tests pro Kanton.

Tabelle 2: Teilnahme nach Kanton

Kantone	Schulen	Klassen	Schülerinnen und Schüler	Anzahl Online-Tests	Anzahl Papier-Tests
Aargau	101	384	6'624	45'265	12'925
Basel-Landschaft	20	144	2'628	20'702	5'207
Basel-Stadt	11	96	1'366	10'590	2'651
Solothurn	35	146	2'387	18'591	4'695
Total	177	770	13'005	95'148	25'478

5 Testdurchführung

5.1 Testzeitfenster

Die Durchführung des Check S2 fand zwischen dem 18. Februar und dem 29. März 2019 statt. Die Ergebnisse wurden am 3. Mai 2019 zurückgemeldet.

5.2 Anzahl durchgeführter Tests

Während des gesamten Zeitfensters nahmen 13'005 Schülerinnen und Schüler aus den Kantonen Aargau, Basel-Landschaft, Basel-Stadt und Solothurn am Check S2 2019 teil. Die Schülerinnen und Schüler lösten in den Fächern Deutsch, Englisch, Französisch, Mathematik sowie Natur und Technik insgesamt 95'148 Online-Tests.[2] Zusätzlich wurden 12'841 Texte in Deutsch und 12'637 Texte in Englisch verfasst und beurteilt.

6 Online-Tests: Auswertung

6.1 Scoring

Das Scoring der Online-Tests, das Übersetzen also der Antworten in numerische Werte, wurde direkt im Onlinesystem (*assessment delivery platform*) vorgenommen. Dabei wurde zwischen richtig gelösten, falsch gelösten und nicht bearbeiteten Aufgaben unterschieden. Das ordnungsgemässe

[2]Die Zahl der Online-Tests kann von weiteren publizierten Zahlen abweichen, da sie lediglich die ausgewerteten und zurückgemeldeten Online-Tests für öffentliche Schulen des Bildungsraums Nordwestschweiz beinhaltet.

Funktionieren des Scorings wurde vor dem Testzeitfenster mehrfach von Projektmitarbeitenden am IBE überprüft. Für offene Aufgabenstellungen (Textfelder) steht dem System ein Set an Antwortalternativen zur Verfügung. Zusätzlich wurde das vom System vorgenommene Scoring der offenen Aufgaben im Rahmen der Auswertung manuell nachgeprüft, um sicherzustellen, dass allfällige bei den Antwortalternativen nicht bedachte (aber richtige) Antworten korrekt beurteilt wurden. Vor der Berechnung der Ergebnisse der Online-Tests wurde die Qualität sämtlicher Aufgaben überprüft (Testgütekriterien, Modellkonformität, Differential Item Functioning (DIF)). Für die Berechnung der Testergebnisse wurden Aufgaben, deren teststatistische Gütekriterien ungenügend waren, ausgeschlossen (siehe Tabelle 3 zur Anzahl Ausschlüsse pro Skala).

Des Weiteren wurde überprüft, wie viele Aufgaben von den Schülerinnen und Schülern tatsächlich bearbeitet wurden. Einige Schülerinnen und Schüler haben einzelne Aufgaben nicht bearbeitet, beispielsweise aus Zeitgründen, aufgrund technischer Probleme oder weil sie die Lösung nicht wussten. Wurde ein Testteil korrekt abgeschlossen, aber nicht alle Aufgaben bearbeitet, dann wurden nicht bearbeitete Aufgaben als falsch gewertet. Ein Testergebnis wurde allerdings nur dann berechnet, wenn eine Schülerin oder ein Schüler mindestens zwei Testteile eines Online-Tests und mindestens 20 Prozent aller Aufgaben bearbeitet hatte. Wurden alle Aufgaben eines Testteils nicht bearbeitet, dann wurde der entsprechende Testteil für die Berechnung der Testergebnisse ausgeschlossen.

Neben der Punktzahl für jeden Kompetenzbereich wurden fachübergreifende Gesamtwerte ausgewiesen. Die Gesamtwerte in Deutsch, Englisch, Französisch und Mathematik wurden als arithmetische Mittelwerte aus den Ergebnissen der einzelnen Kompetenzbereiche berechnet. In Natur und Technik wird jeweils nur ein Gesamtwert zurückgemeldet. Generell wurden die Gesamtwerte nur dann berechnet, wenn für die jeweiligen Kompetenzbereiche allesamt gültige Testergebnisse vorlagen.

6.2 Skalierung und Parameterschätzung

Die Skalierung der Daten aus den Online-Tests erfolgte mit dem Softwarepaket TAM von Robitzsch, Kiefer und Wu (2019) in der Entwicklungsumgebung R (R Core Team, 2019). Für die Auswertung der dichotomen Aufgaben beziehungsweise Items kam ein Zwei-Parameter-Logistisches-Modell (auch Birnbaum-Modell; vgl. Birnbaum, 1958) zum Einsatz, bei dem neben der Personenfähigkeit (θ) und der Itemschwierigkeit (β) auch die Itemtrennschärfe (α) geschätzt wird (vgl. De Ayala, 2009). Damit bestimmt sich die Lösungswahrscheinlichkeit als

$$p\left(x_j = 1 | \theta_i, \alpha_j, \beta_j\right) = \frac{e^{\alpha_j(\theta_i - \beta_j)}}{1 + e^{\alpha_j(\theta_i - \beta_j)}},$$

sodass die Items eine unterschiedliche Position auf der latenten Fähigkeits-Schwierigkeits-Dimension und eine unterschiedliche Steigung einnehmen können. Es wird zudem von einem zugrunde liegenden Populationsmodell ausgegangen, das die Personenparameter (θ) als normalverteilt annimmt. Gegenüber dem einfacheren Ein-Parameter-Modell (auch Rasch-Modell; vgl. Rasch, 1960), bei dem die Steigungsparameter über alle Items gleichgesetzt werden, ergibt sich beim Zwei-Parameter-Modell in der Regel der Vorteil einer besseren Passung zu den Daten. Im Zwei-Parameter-Modell werden

die Items unterschiedlich stark, nämlich proportional zu ihrer Trennschärfe, für die Berechnung der Fähigkeiten berücksichtigt. Somit werden trennscharfe Items für die Fähigkeitsschätzung stärker gewichtet als wenig trennscharfe Items (vgl. Birnbaum, 1968).

Die Kalibrierung der Aufgaben wurde basierend auf Tests von Schülerinnen und Schülern durchgeführt, welche von ihnen zu einem vorher definierten Stichdatum (meist kurz vor dem Ende des Testzeitfensters) bearbeitet worden waren. Es wurde darauf geachtet, dass der Kalibrierungsdatensatz den grössten Teil (in der Regel $>$ 90 Prozent) der Tests der teilnehmenden Schülerinnen und Schüler beinhaltete. In diesem Kalibrierungsdatensatz wurden alle Schülerinnen und Schüler ausgeschlossen, die mehr als 20 Prozent fehlende Werte auf den Items des jeweiligen Tests hatten. Dieser Ausschluss betraf nur einen kleinen Teil der Schülerinnen und Schüler (in der Regel $<$ 5 Prozent) und soll die Stabilität der Parameterschätzungen erhöhen.

Die Kalibrierung erfolgte in einem oder mehreren Schritten, in denen in der Regel Items mit geringer CTT-Trennschärfe von $r_{tt} < .20$, berechnet als punkt-biseriale Korrelation zwischen dem Einzelitem und dem Gesamttest, mit extremer Schwierigkeit (in der Regel $P < .05$ und $P > .95$) oder mit einer auffällig hohen oder tiefen Infit- bzw. Outfit-Statistik (in der Regel Infit/Outfit $< .70$ und Infit/Outfit > 1.30; vgl. Wright & Linacre, 1994)[3] ausgeschlossen und für die weiteren Auswertungen nicht berücksichtigt wurden. Zusätzlich wurden Items ausgeschlossen, deren modellbasierte Item-Charakteristik-Kurven (ICC) nicht zu den beobachteten Lösungswahrscheinlichkeiten passten. Aufgrund inhaltlicher Überlegungen wurden einzelne Aufgaben, die diesen Item-Fit-Statistiken nicht vollständig genügten, dennoch beibehalten (z.B. «Eisbrecheritems»). Der Ausschluss von Items basiert dementsprechend immer auf i) empirischen Item-Fit-Statistiken im Zusammenhang mit ii) inhaltlichen Überlegungen. Pro Test mussten nur wenige Aufgaben ausgeschlossen werden, wie Tabelle 3 zeigt.

6.3 Testlinking

Damit individuelle Lernzuwächse sowie Unterschiede in den geprüften fachspezifischen Kompetenzen im Kohortenvergleich auf einer einheitlichen und kontinuierlichen Skala abgebildet werden können, wurde ein Linking der Tests über unterschiedliche Testdurchführungen hinweg vorgenommen. Das Testdesign entspricht dem *common-item non-equivalent groups design* (vgl. Kolen & Brennan, 2004), wobei in den jährlich unterschiedlichen Tests jeweils ein Teil gleicher Aufgaben wiederholt eingesetzt wird. Die Itemparameter bereits verwendeter Aufgaben wurden dementsprechend zur Skalenstabilisierung fixiert, das heisst auf den bisher geschätzten Itemparameterwerten belassen (Link-Items). Dies erlaubte es, die Schwierigkeit und die Trennschärfe der neuen Aufgaben auf der bestehenden Skala einzuordnen. Die Link-Items ermöglichten es somit, die unterschiedlichen Tests miteinander auf einer einheitlichen Schwierigkeitsskala zu vergleichen und gesamtheitliche Aussagen über fachspezifische Kompetenzen im Kohortenvergleich zu treffen. Die Personenparameter (Fähigkeiten) wurden mittels WLE (*Weighted Likelihood Estimation*; vgl. Warm, 1989) geschätzt. Die Metrik der geschätzten Item- und Personenparameter basierend auf IRT-Modellen ist grundsätzlich arbiträr und die

[3] Diese Regel kann aufgrund ihrer Strenge in der Praxis nicht immer eingehalten werden.

Tabelle 3: Item-Ausschluss, Link-Items und EAP-Reliabilität pro Skala

Skala	Anzahl Items Total	Anzahl Ausschlüsse	Anzahl Link-Items	EAP-Reliabilität
Mathematik	231	2	185	0.93
Deutsch				
Lesen	200	6	84	0.90
Sprache im Fokus	170	8	58	0.87
Französisch				
Lesen	150	9	52	0.83
Hören	150	8	71	0.82
Englisch				
Lesen	160	5	63	0.93
Hören	150	5	61	0.89
Natur und Technik	196	10	59	0.85

Parameter können linear auf eine beliebige Skala, z.B. auf die Check-Skala von 0 bis 1200 Punkten, transformiert werden.

Im Frühling 2019 wurde die Check-Skala angepasst. Neu werden die Ergebnisse sämtlicher Checks und von Mindsteps auf einer gemeinsamen Skala zurückgemeldet. Damit lässt sich der Lernfortschritt der Schülerinnen und Schüler von der 3. Klasse der Primarstufe bis zur 3. Klasse der Sekundarstufe I abbilden. Als Referenz für die neue Skala dienten der Check S2 2019 und der Check P6 2018. Für den Check S2 2019 wurde der Mittelwert der Schülerinnen und Schüler neu auf 800 Punkte standardisiert. Die Standardabweichung wurde in Abhängigkeit zum Lernfortschritt zwischen dem Check P6 2018 (Mittelwert 650 Punkte) und dem Check S2 2019 fixiert. Die Ergebnisse des diesjährigen Check S2 können deshalb nicht mit den Ergebnissen der Vorjahre verglichen werden.

Zur Sicherstellung der Qualität des Testlinkings wurden die Link-Items auf Differential Item Functioning (DIF) überprüft. Das heisst, dass überprüft wurde, ob die Link-Items in unterschiedlichen Schülerjahrgängen gleich funktionieren oder ob für diese Aufgaben bei gleichen Schülerfähigkeiten unterschiedliche Lösungswahrscheinlichkeiten vorliegen. Beispielsweise könnte es vorkommen, dass in einem Schülerjahrgang über ein verstärktes Üben bestimmter Aufgabenformate/Inhalte gewisse Aufgabenstellungen von schwächeren Schülerinnen und Schülern mit vergleichsweise höherer Wahrscheinlichkeit richtig gelöst werden (oder umgekehrt). Aufgaben mit solchen Verschiebungen der Parameterwerte über die Schülerjahrgänge hinweg eignen sich nicht als Link-Items und werden daher für den betreffenden Jahrgang jeweils neu geschätzt. Die Überprüfung von DIF erfolgte anhand grafischer Vergleiche der geschätzten Item-Charakteristik-Kurven (ICC) und mittels der em-

pirisch beobachteten Lösungswahrscheinlichkeiten nach Personenfähigkeitskategorien. Zudem wurde die RMSD (*Root Mean Square Deviation*)-Statistik als Index zur Bestimmung von DIF verwendet (Regel < 0.12; vgl. OECD, 2017, S. 151). Tabelle 3 zeigt in der letzten Spalte die Anzahl Aufgaben, die im vorliegenden Check pro Skala als Link-Items mit fixierten Parametern verwendet wurden, und die EAP-Reliabilitäten. Im Anhang werden pro Skala die Testinformationen nach Hauptpfaden sowie die Verteilungen der Item- und Personenparameter ausgewiesen. Ebenfalls findet sich im Anhang eine Tabelle der Mittelwerte und Standardabweichungen pro Skala sowie der Skalen-Interkorrelationen.

7 Papier-Tests: Auswertung

7.1 Kriterien zur Beurteilung der Texte

Die Texte in Deutsch Schreiben und Englisch Schreiben wurden von einem Team aus Linguistinnen, Linguisten und Lehrpersonen im entsprechenden Fach beurteilt. Um die Texte der Schülerinnen und Schüler bewerten zu können, wurde ein standardisiertes Beurteilungsraster eingesetzt. Das Beurteilungsverfahren entspricht einem analytischen Verfahren (*analytical scoring*) mit Kriterienraster, bei dem verschiedene Aspekte eines Textes nach verbal formulierten Abstufungen bewertet werden (Weigle, 2002). Die Beurteilung der Texte bezieht sich auf die kommunikativen und linguistischen Fähigkeiten, die sich im Schreibprodukt zeigen (Nussbaumer & Sieber, 1994). Die Beurteilung der Deutschtexte umfasst vier Dimensionen:

1. Inhalt: Auftragserfüllung und Aussagekraft
2. Textaufbau und Textzusammenhang
3. Sprachrichtigkeit
4. Sprachangemessenheit, Schreibstil und Ästhetik

Diese vier Dimensionen wurden mit insgesamt 15 Beurteilungskriterien operationalisiert. Für jedes Kriterium wurden drei oder vier Abstufungen unterschieden. Die Beurteilung der Englischtexte umfasst zwei Dimensionen:

1. Inhalt
2. Sprachrichtigkeit

Diese beiden Dimensionen wurden mit insgesamt acht bis neun Beurteilungskriterien operationalisiert. Für jedes Kriterium wurden drei bis fünf Abstufungen unterschieden.

7.2 Beurteilungsprozess und Qualitätssicherung

Damit alle Beurteiler und Beurteilerinnen die Beurteilungskriterien über die gesamte Korrekturzeit gleich anwenden, wurden im Anschluss an eine zweitägige Schulungsphase täglich fünf bis zehn zufällig ausgewählte Texte von allen beurteilenden Personen bewertet (*Multiple Ratings*) und die Beurteilungen miteinander verglichen. Dies diente zum einen der stetigen Überprüfung des gemeinsamen

Verständnisses der Bewertungskriterien und zum anderen als direktes Feedback an die einzelnen Beurteiler und Beurteilerinnen hinsichtlich ihrer Positionierung auf dem Strenge-Milde-Massstab. Zusätzlich wurden pro Schulfach jeweils etwa 120 Texte doppelt korrigiert (*double ratings*). Mit dieser Vorgehensweise wurde ein einheitlicher Beurteilungsmassstab und damit eine hohe Beurteilungskonsistenz angestrebt. Dass dies gelungen ist, zeigen die Auswertungen der Multiple Ratings in Tabelle 4 für Deutsch und in Tabelle 5 für Englisch. Bestimmt wurde die Intraklassenkorrelation r_{ICC} (genauer: die *oneway multiple raters consistency* nach McGraw & Wong, 1996) für jede einzelne Beurteilungsdimension ohne Berücksichtigung des Aufsatzthemas. Koo und Li (2016) schlagen vor, Übereinstimmungen grösser als .50 als «ausreichend», grösser als .75 als «gut» und grösser als .90 als «hervorragend» zu bezeichnen und sich dabei an dem 95%-Konfidenzintervall der Punktschätzung zu orientieren. Demnach ist bei jeder einzelnen Skala die Übereinstimmung als mehr als hervorragend zu bezeichnen. Mit an Sicherheit grenzender Wahrscheinlichkeit liegt der r_{ICC} für alle Skalen über .90, dies ohne augenfällige Unterschiede zwischen den Fächern Deutsch und Englisch.

Tabelle 4: Beurteilerübereinstimmung Multiple Ratings Deutsch Schreiben

Skala	$N_{subjects}$	N_{rater}	r_{ICC}	$\text{CI}_{95\%}$	$\text{P}(r > .90)$
Inhalt	136	10	.92	[.90, .94]	<.05
Textaufbau	136	10	.93	[.91, .95]	<.05
Sprachrichtigkeit	136	10	.97	[.97, .98]	<.001
Sprachangemessenheit	136	10	.93	[.91, .95]	<.001

Tabelle 5: Beurteilerübereinstimmung Multiple Ratings Englisch Schreiben

Skala	$N_{subjects}$	N_{rater}	r_{ICC}	$\text{CI}_{95\%}$	$\text{P}(r > .90)$
Aufsatz 1					
Inhalt	96	9	.97	[.96, .98]	<.001
Sprachrichtigkeit	96	9	.97	[.96, .98]	<.001
Aufsatz, einfache Version					
Inhalt	46	9	.96	[.94, .97]	<.001
Sprachrichtigkeit	45	9	.97	[.96, .98]	<.001
Aufsatz, schwierige Version					
Inhalt	49	9	.96	[.94, .97]	<.001
Sprachrichtigkeit	49	9	.97	[.96, .98]	<.001

7.3 Skalierung und Parameterschätzung

Die Skalierung der Papier-Tests erfolgte mit der Software ConQuest (Wu, Adams, Wilson & Haldane, 2007) auf der Basis eines Multifacetten-Rasch-Modells. In diesem Modell kann berücksichtigt werden, dass derselbe Text von unterschiedlichen Personen – trotz vorgegebener Kriterien, Schulungsphase und generell guter Beurteilerübereinstimmung – dennoch nicht immer genau gleich streng beurteilt wird. Dies lässt sich aufgrund des Interpretationsspielraums bei offen gestellten Aufgaben nicht verhindern. Während gebundene Testaufgaben eindeutig als richtig oder falsch korrigiert werden können, spielt der Beurteilungsmassstab der beurteilenden Personen (Rater) bei offenen Aufgabenformaten eine Rolle für das Testergebnis. Beurteilt beispielsweise Rater A systematisch strenger als Rater B, dann ist dies für all jene Schülerinnen und Schüler ungerecht, deren Texte von Rater A beurteilt werden. Wird die Strenge oder Milde in der Beurteilung der Texte bei der Berechnung der Testergebnisse nicht berücksichtigt, dann wird ein Text je nach der beurteilenden Person entweder besser oder weniger gut beurteilt. Aus diesem Grund wurde die Beurteilungsstrenge der beurteilenden Personen als Facette der Urteilssituation aufgefasst und bei der Berechnung der Ergebnisse wie folgt berücksichtigt:[4]

$$\ln\left(\frac{P_{ijnm}}{P_{ijn(m-1)}}\right) = \theta_i - (R_n + \beta_j + F_m)$$

P_{ijnm} und $P_{ijn(m-1)}$ entsprechen der Wahrscheinlichkeit, dass die Person i von Rater n die Beurteilung m beziehungsweise $m-1$ erhält. θ_i entspricht der Fähigkeit der Person i, R_n der Strenge des Raters n, β_j der Schwierigkeit des Items j und F_m der Schwierigkeit des Beurteilungsschritts m relativ zum Beurteilungsschritt $m-1$ (*rating scale steps*). Im Rahmen der Kalibrierung wurden die CTT-Trennschärfe und die CTT-Schwierigkeit, die Infit- und Outfit-Statistiken sowie die ICC der Beurteilungskriterien geprüft. Die Multifacetten-Analyse lieferte zudem für jede beurteilende Person eine Schätzung für die Beurteilungsstrenge auf der Logit-Skala, einen dazugehörigen Standardfehler (Genauigkeit der Schätzung der Strenge) sowie Informationen zur Modellkonformität der Schätzung der Beurteilungsstrenge (Infit- und Outfit-Statistiken). Im Gegensatz zu den Online-Tests werden die Papier-Tests nicht über die Jahre hinweg gelinkt. Das Beurteilungsraster wird zwar jeweils zwischen zwei Erhebungen nicht oder nur minim angepasst. Allerdings werden jährlich neue Themen definiert und neue Rater rekrutiert und ausgebildet. Deshalb können die Ergebnisse in Deutsch und Englisch Schreiben nicht direkt über die Jahre hinweg verglichen werden. Die Personenparameter (Fähigkeiten) wurden – analog zu den Online-Tests – mittels WLE (*weighted likelihood estimation*, vgl. Warm, 1989) geschätzt und auf die Check-Skala transformiert. Der Mittelwert der Schülerinnen und Schüler wurde dabei für Deutsch Schreiben auf den Mittelwert der beiden Kompetenzbereiche Deutsch Lesen und Deutsch Sprache im Fokus standardisiert. Der Mittelwert in Englisch Schreiben wurde auf den Mittelwert der beiden Kompetenzbereiche Englisch Lesen und Englisch Hören standardisiert. Die Standardabweichung wurde für beide Papier-Tests auf 100 Punkte festgelegt.

[4] Bei der Skalierung wurde nicht berücksichtigt, welches Thema die Schülerinnen und Schüler gewählt hatten, da nicht ausgeschlossen werden kann, dass ein Thema nur deshalb schwieriger erscheint, weil es besonders häufig von schwächeren Schülerinnen und Schülern gewählt wurde.

Durch den Einbezug der Rater in das Modell können gleiche Beurteilungen von unterschiedlichen Ratern in leicht unterschiedlichen Fähigkeitsschätzungen und damit auch in leicht unterschiedlichen Punktzahlen auf der Check-Skala resultieren.

8 Glossar

Das Glossar ist thematisch aufgebaut.

Item-Response-Theorie (IRT). Ist die probabilistische Testtheorie, die die Grundlage für unterschiedliche statistische Modelle zur Testauswertung bildet. Sie baut auf der Annahme auf, dass die über einen Test zu messenden Eigenschaften latent, d.h. nicht direkt beobachtbar sind. Aus dem Antwortverhalten auf Testaufgaben können Rückschlüsse auf die zu messenden, latenten Eigenschaften gezogen bzw. Zusammenhänge formuliert werden. Es wird davon ausgegangen, dass das Antwortverhalten in Abhängigkeit von Personen- sowie Aufgabeneigenschaften hervorgebracht wird.

Klassische Testtheorie (CTT). Beschreibt eine fachgeschichtlich ältere Testtheorie, deren Kritikpunkte die IRT zu überwinden sucht. Eine wesentliche Annahme der Klassischen Testtheorie ist, dass sich die anhand eines Tests ermittelte Eigenschaft (z.B. fachspezifische Kompetenz) einer Person aus dem «wahren Wert» der Person und einem testspezifischen Messfehler zusammensetzt.

Personenfähigkeit (θ). Kennwert der Person. Gibt auf einer metrischen Skala an, wie gut die Person im Vergleich zu anderen Personen aus der Population die zu messende Eigenschaft/Fähigkeit besitzt. Die gewählte Metrik ist willkürlich (Standard: zentriert um den Nullpunkt mit einer Standardabweichung von 1). Die Personenfähigkeiten können linear in eine andere Skala (Punktzahlen innerhalb eines Referenzrahmens, z.B. Check-Skala) transformiert und zurückgemeldet werden.

Itemschwierigkeit (β). Kennwert der Aufgaben nach probabilistischer Testtheorie. Wird auf der gleichen Skala wie die Personenfähigkeit abgebildet. Gibt die Personenfähigkeit an, die benötigt wird, um die entsprechende Aufgabe mit einer Wahrscheinlichkeit von 50 Prozent lösen zu können.

Trennschärfe (α). Kennwert der Aufgaben nach probabilistischer Testtheorie. Die Trennschärfe beschreibt, wie gut die Aufgabe zwischen fähigeren und weniger fähigen Personen differenziert. Eine hohe Trennschärfe bedeutet, dass fähigere Personen eine deutlich höhere Wahrscheinlichkeit haben, die betreffende Aufgabe richtig zu lösen, als weniger fähige Personen. Eine niedrige Trennschärfe (nahe null) bedeutet, dass sich die Lösungswahrscheinlichkeiten der entsprechenden Aufgabe zwischen fähigen und weniger fähigen Personen nicht stark unterscheiden.

CTT-Schwierigkeit. Kennwert der Aufgaben nach Klassischer Testtheorie. Gibt den Anteil der Personen an, die die Aufgabe korrekt gelöst haben.

CTT-Trennschärfe (r_{it}). Kennwert der Aufgaben nach Klassischer Testtheorie. Gibt die Korrelation der betreffenden Testaufgabe mit dem Gesamtscore an. Eine hohe Trennschärfe nach Klassischer Testtheorie bedeutet, dass das einzelne Item zwischen Personen mit hoher bzw. niedriger Kompetenz im Sinne des Gesamttests differenziert.

Itemanalyse. Überprüfen der Eignung der verwendeten Aufgaben (Item-Fit) zur Messung der gewünschten Eigenschaft (z.B. fachspezifische Kompetenz) anhand unterschiedlicher statistischer Verfahren, Gütekriterien und Visualisierungen.

Kalibrierung. Bezeichnet die Schätzung der Itemparameter (z.B. Schwierigkeiten und Trennschärfen) basierend auf einem zugrunde liegenden IRT-Modell.

Linking. Verorten von Aufgaben (Itemschwierigkeiten) auf einer einheitlichen Skala (Referenzrahmen) über verschiedene Testdurchführungen hinweg. Ermöglicht es, Fortschritte auf einer einheitlichen und kontinuierlichen Skala abzubilden.

Skalierung. Die Begriffe Skalierung und Kalibrierung werden zum Teil synonym verwendet. Je nach Literatur wird der Begriff Skalierung noch etwas umfassender verwendet und bezieht sich auf den Gesamtprozess der Itemanalysen sowie die (simultane) Itemparameter- und Personenparameterschätzung.

EAP-Reliabilität (ρ_{EAP}). Expected A Posteriori Estimation Reliability Testgütekriterium. Dient der Überprüfung, ob die verwendeten Aufgaben in einem Test Unsicherheit bei der Verortung der getesteten Personen auf dem latenten Konstrukt zu reduzieren vermögen. Kann Werte zwischen 0 (je nachdem auch negativ) und 1 annehmen, wobei Werte nahe 1 für hohe Messpräzision sprechen.

Testinformation. Im IRT-Kontext variiert die Messgenauigkeit mit der Ausprägung der Personenfähigkeit. Die Verteilung der Testinformation gibt grafisch den Zusammenhang zwischen der statistischen Information in den Daten (y-Achse) und den geschätzten Personenfähigkeiten (x-Achse) wider. Der Gipfel der Kurve gibt an, welche Bandbreite an Personenfähigkeiten der Test am zuverlässigsten misst. Messungen in Extrembereichen der Fähigkeitsskala sind weniger genau; hier nimmt die Testinformation ab bzw. der Standardfehler des Tests zu.

Item-Charakteristik-Kurven (ICC). Grafische Darstellung des Zusammenhangs zwischen Lösungswahrscheinlichkeit (y-Achse) und Personenfähigkeit (x-Achse) pro Testaufgabe. Für das verwendete Birnbaum-Modell kann dieser Zusammenhang anhand einer S-förmigen Kurve skizziert werden. Mit steigender Personenfähigkeit sollte erwartungsgemäss die Lösungswahrscheinlichkeit zunehmen.

Differential Item Functioning (DIF). Aufgaben, die für verschiedene Personengruppen trotz gleicher Personenfähigkeiten unterschiedlich schwierig zu lösen sind. Das heisst, dass neben der zu messenden Personenfähigkeit weitere Faktoren die Lösungswahrscheinlichkeiten der entsprechenden Aufgaben beeinflussen. Die Abwesenheit von DIF wird im Rahmen der Itemanalyse geprüft.

Root Mean Square Deviation (RMSD). Die RMSD-Statistik quantifiziert in Form eines standardisierten Indexes die Diskrepanz zwischen der beobachteten ICC und der erwarteten ICC. Werte < 0.05 sprechen für eine gute Passung. Werte ≥ 0.12 weisen auf Abweichung zwischen der beobachteten und der erwarteten ICC und somit auf DIF hin.

Outfit. Residuenbasierte Item-Fit-Statistik. Ausmass an Passung zwischen der Steigung der ICC basierend auf den beobachteten Antworten und der gemäss Modell erwarteten Steigung der ICC. Die Outfit-Statistik hat einen Erwartungswert von 1. Bewegt sich die Outfit-Statistik zwischen 0.70 und 1.30, passen die Daten zum Modell.

Infit. Residuenbasierte Item-Fit-Statistik. Gewichtete Version der Outfit-Statistik, wobei Ausreisser (d.h. Personen mit geschätzten Fähigkeiten, für die das Item weniger informativ ist) weniger stark gewichtet werden. Die Infit-Statistik hat einen Erwartungswert von 1. Liegt die Infit-Statistik zwischen 0.70 und 1.30, passen die Daten zum Modell.

Rater-Effekte. (Ungewollte) Einflüsse auf die Testergebnisse durch unterschiedliche Beurteilungsstrenge der beurteilenden Personen bei der Korrektur von offenen Aufgaben. Rater-Effekte können im Rahmen von Multifacetten-Modellen kontrolliert werden.

Intraklassenkorrelation (r_{ICC}). Dient als Reliabilitätsmass und gibt das Ausmass an Übereinstimmung zwischen den Beurteilungen unterschiedlicher Rater an. Werte nahe 1 weisen darauf hin, dass zwischen den Beurteilungen der Rater keine grossen Unterschiede bestehen. Werte nahe 0 bedeuten, dass sich die Rater in ihren Einschätzungsergebnissen stark unterscheiden.

Standardabweichung. Die mittlere Abweichung der einzelnen Werte zum Mittelwert.

Standardfehler. Mass der Messgenauigkeit. Je grösser der Standardfehler, desto ungenauer ist die Schätzung. Der Bereich +/– zweimal den Standardfehler (Konfidenzintervall) sollte mit 95-prozentiger Wahrscheinlichkeit den wahren Wert der getesteten Person auf der Fähigkeitsskala enthalten.

9 Literaturverzeichnis

Birnbaum, A. (1958). *On the estimation of mental ability (Series Report No. 15)*. Randolph Air Force Base: USAF School of Aviation Medicine.

Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S. 425–435). Reading: Addison-Wesley.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guildford Press.

Duanli, Y., von Davier, A. A. & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. London: Chapman & Hall/CRC.

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

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients. *Journal of Chiropractic Medicine*, *15*, 155–163.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Nussbaumer, M. & Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In P. Sieber (Hrsg.), *Sprachfähigkeiten – besser als ihr Ruf und nötiger denn je! Ergebnisse aus einem Forschungsprojekt* (S. 141–186). Aarau: Sauerländer.

OECD. (2017). *PISA 2015 Technical Report*. OECD: Paris.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. Zugriff 6. September 2019 unter https://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Danmarks Paedagogiske Institut.

Robitzsch, A., Kiefer, T. & Wu, M. (2019). *TAM: Test analysis modules*. R package version 3.1-45. Zugriff 6. September 2019 unter https://CRAN.R-project.org/package=TAM.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370–371.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software manual*. Melbourne: Australian Council for Educational Research.

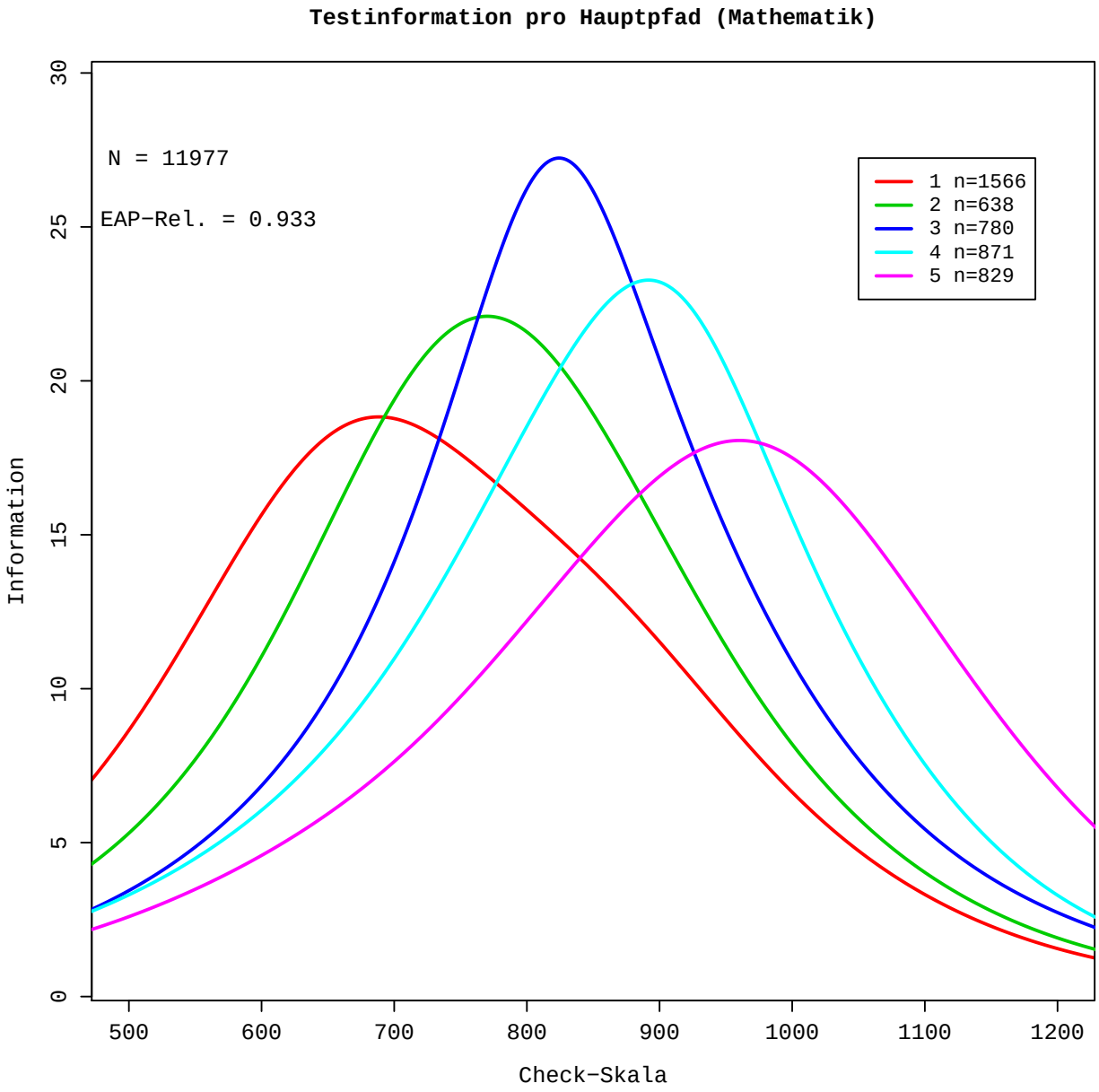
10 Anhang

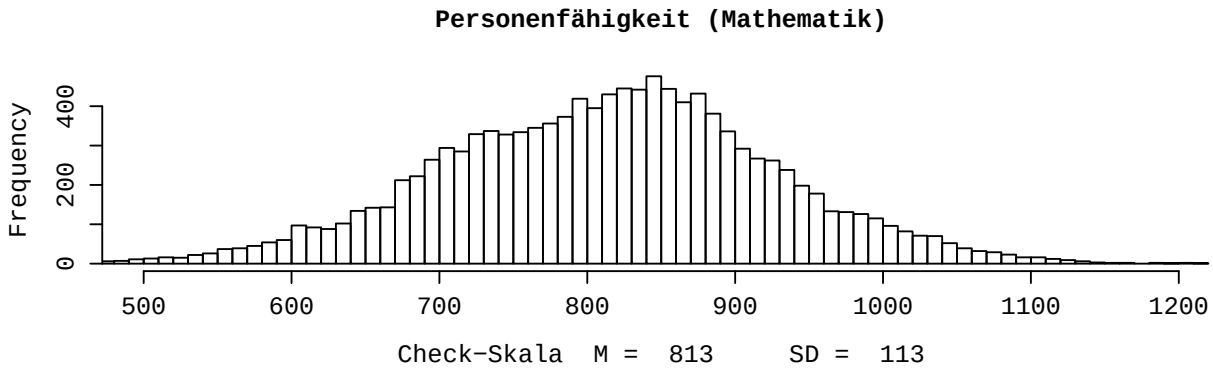
Im Anhang wird pro Skala (Online-Tests) die Testinformation nach Hauptpfad (1–5) grafisch dargestellt. Hauptpfade beschreiben hier ein vertikales Durchlaufen des Tests, wobei Schülerinnen und Schüler die Testteile 2 bis 4 auf derselben Schwierigkeitsstufe lösen (vgl. Abbildung 1). Hauptpfad 1 entspricht z.B. einem Pfad, bei dem Schülerinnen und Schüler die Testteile 2 bis 4 auf der leichtesten Schwierigkeitsstufe lösen. Hauptpfad 5 entspricht demgegenüber einem Pfad, bei dem die Testteile 2 bis 4 auf der höchsten Schwierigkeitsstufe gelöst wurden. Zudem werden pro Skala Verteilungen der Item- und Personenparameter sowie Infit- und Outfit-Statistiken der Kalibrierungsstichproben ausgewiesen. Da Schülerinnen und Schüler mit mehr als zwanzig Prozent an fehlenden Werten für die Kalibrierung ausgeschlossen wurden, unterscheiden sich die berichteten Mittelwerte und Standardabweichungen der Personenparameter von jenen der Gesamtpopulation, welche im Anhang in Tabelle 6 dargestellt werden.

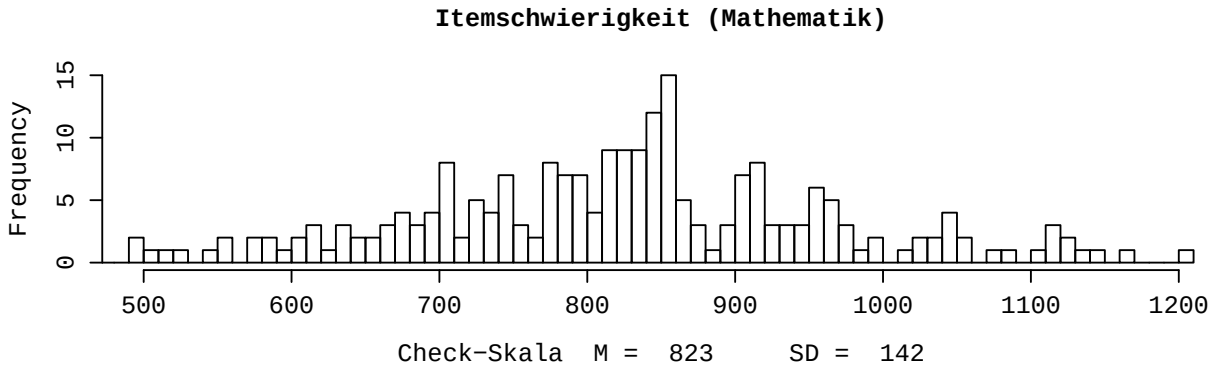
Der Online-Test in Mathematik umfasst drei Subskalen. Diese werden hier gemeinsam dargestellt. Für die Transformation der Personenfähigkeiten und Itemschwierigkeiten auf die Check-Skala wurden die Mittelwerte und Standardabweichungen der drei Subskalen gemittelt.

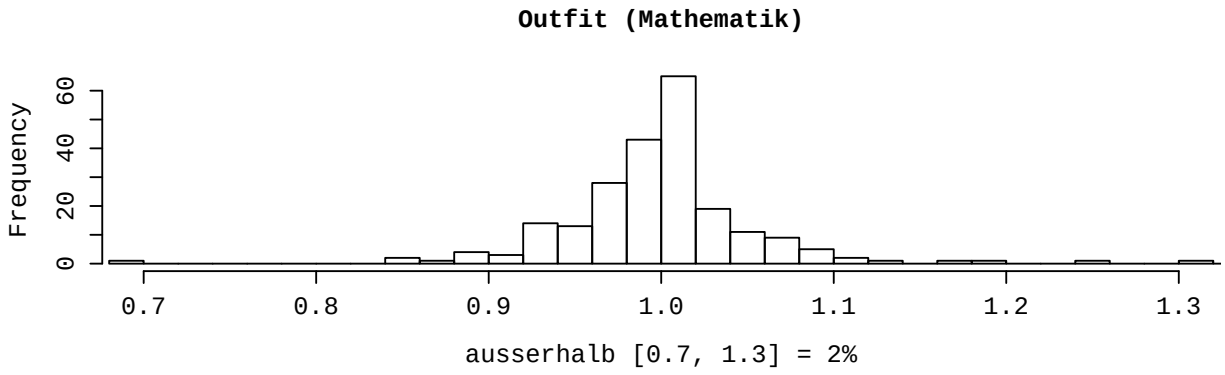
Für den Online-Test Natur und Technik wird die Testinformation lediglich für den obligatorischen Testteil «Allgemeine Naturwissenschaften» angegeben. Da dieser Test statt fünf nur drei Schwierigkeitsstufen enthält, können nur drei Hauptpfade dargestellt werden.

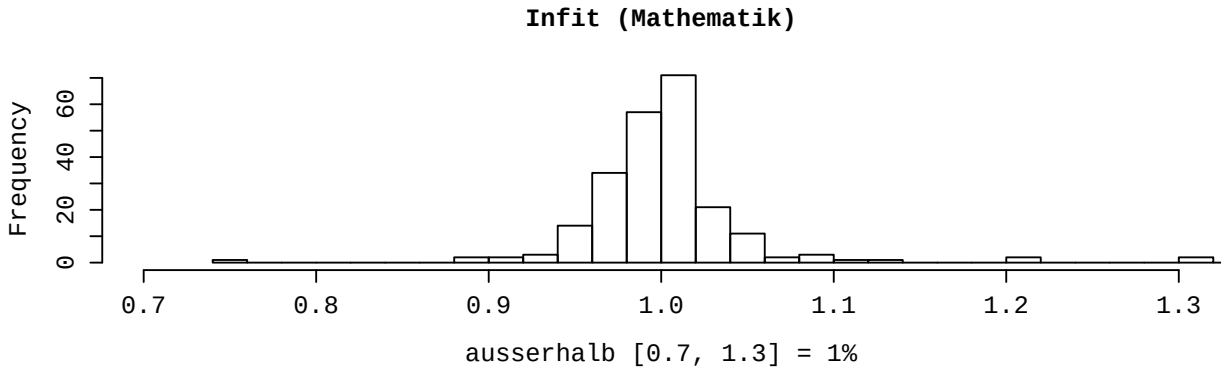
Daneben werden im Anhang Tabellen mit Skalenmittelwerten und -standardabweichungen auf der Check-Skala sowie Skalen-Interkorrelationen ausgewiesen.

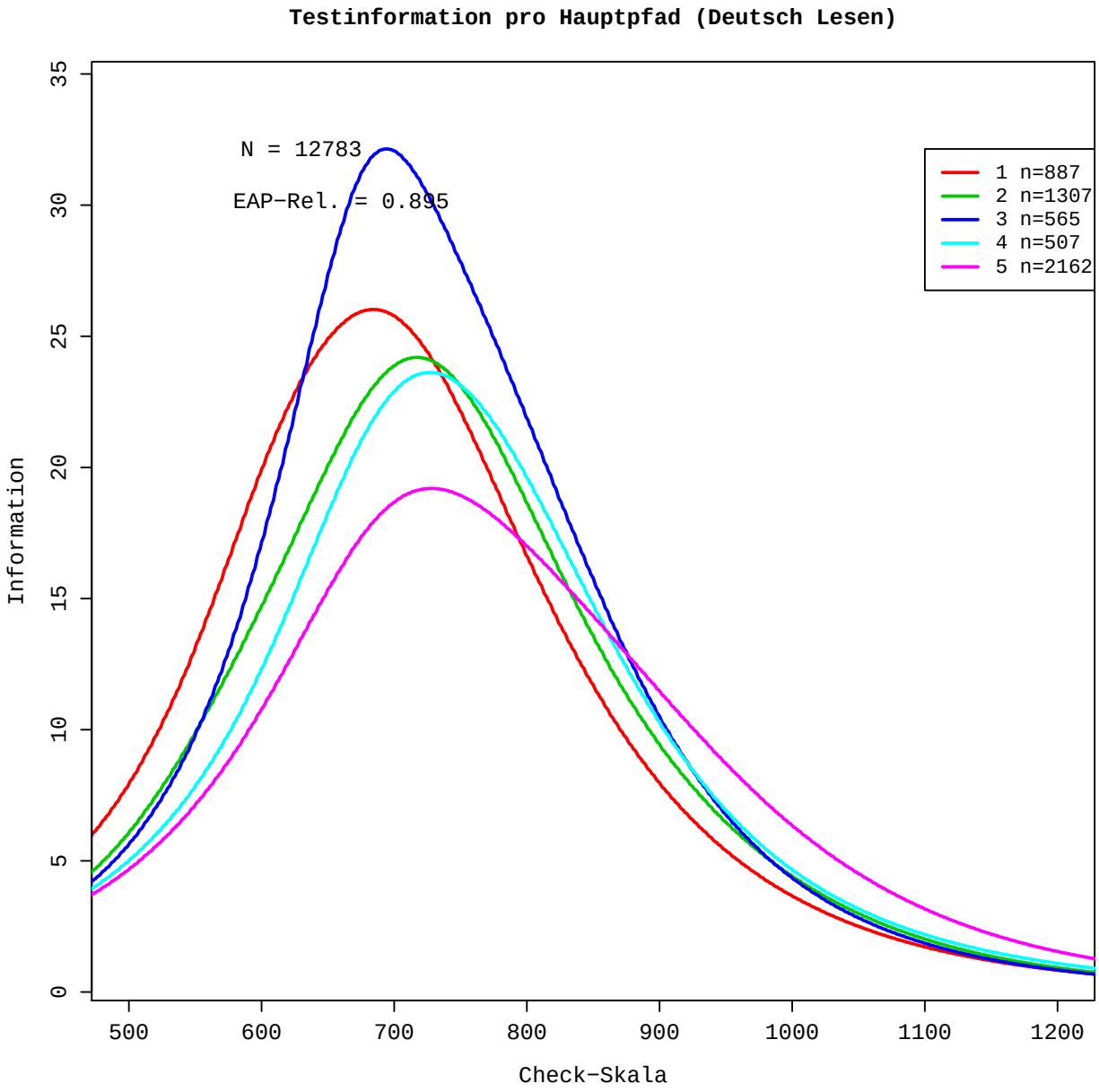

Testinformation pro Hauptpfad (Mathematik)
N = 11977
EAP−Rel. = 0.933
1 n=1566
2 n=638
3 n=780
4 n=871
5 n=829
Information
Check−Skala
0
5
10
15
20
25
30
500
600
700
800
900
1000
1100
1200

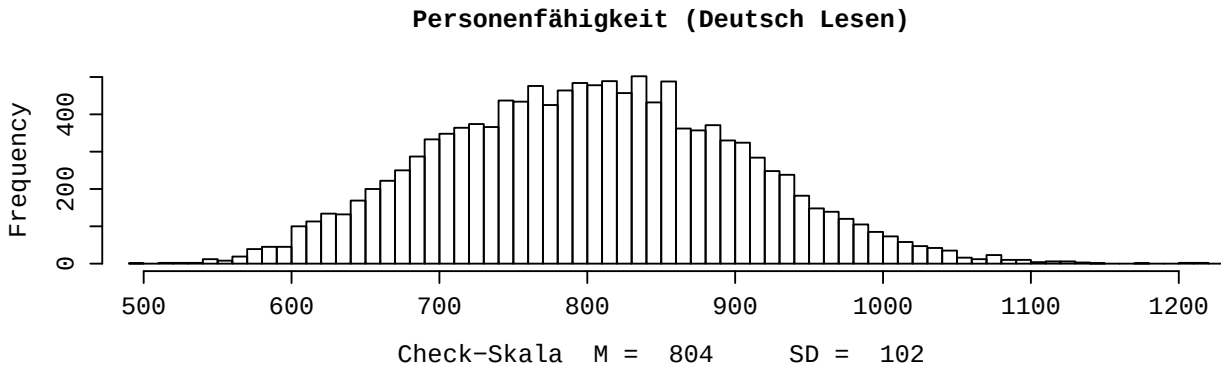

Personenfähigkeit (Mathematik)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 813 SD = 113

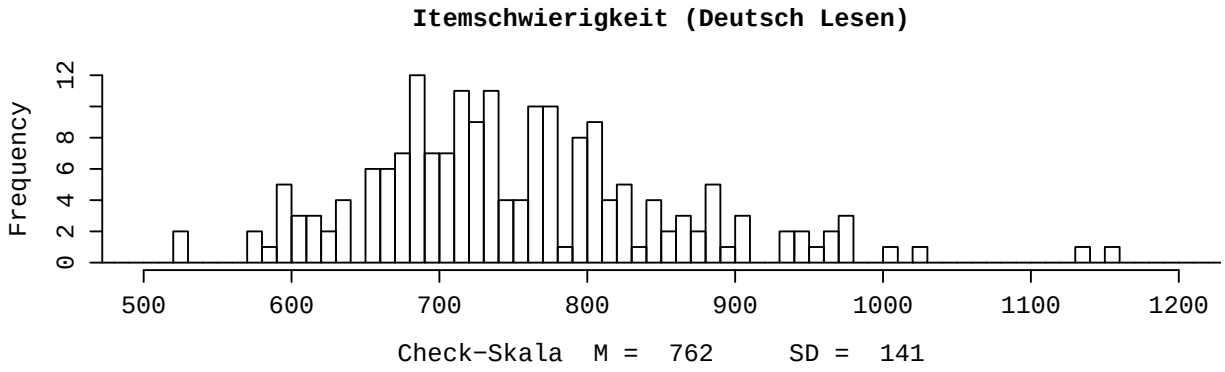

Itemschwierigkeit (Mathematik)
Frequency
0
5
10
15
500
600
700
800
900
1000
1100
1200
Check−Skala M = 823 SD = 142

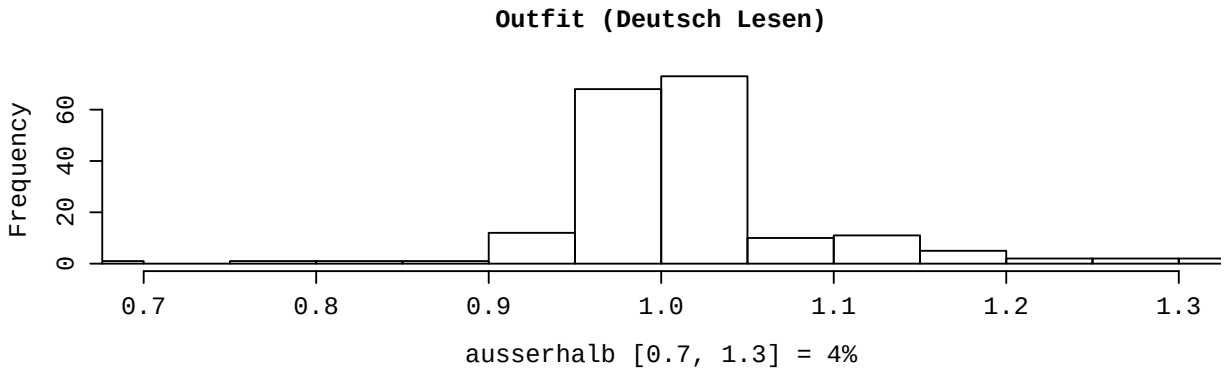

Outfit (Mathematik)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 2%

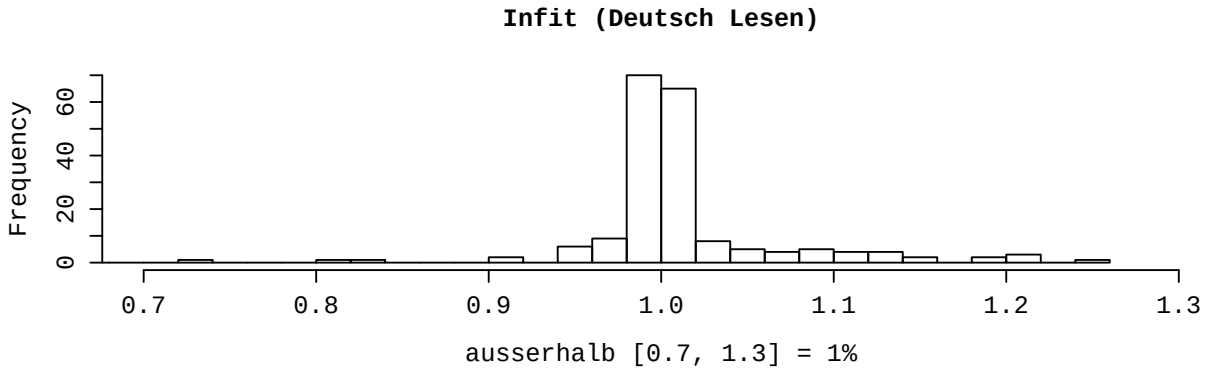

Infit (Mathematik)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

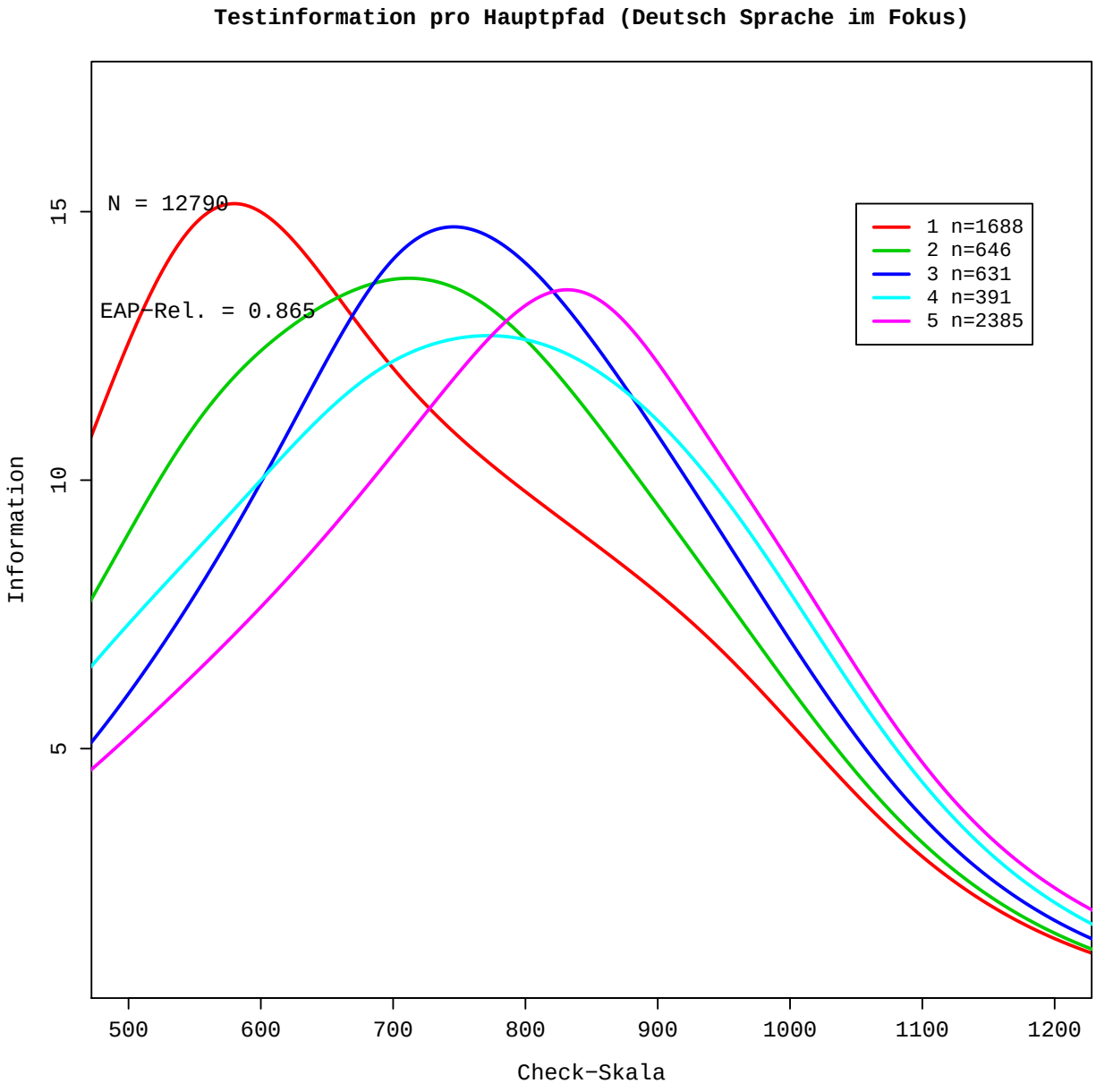

Testinformation pro Hauptpfad (Deutsch Lesen)
N = 12783
EAP−Rel. = 0.895
1 n=887
2 n=1307
3 n=565
4 n=507
5 n=2162
Information
0
5
10
15
20
25
30
35
500
600
700
800
900
1000
1100
1200
Check−Skala

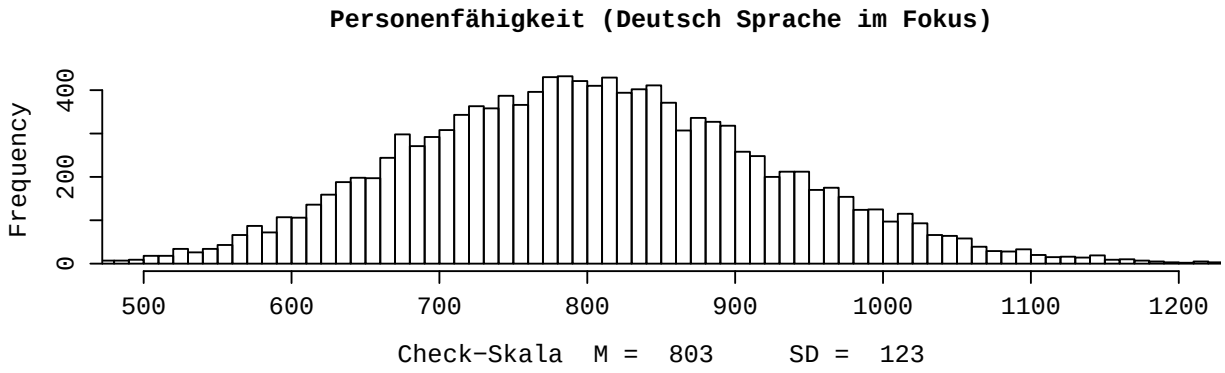

Personenfähigkeit (Deutsch Lesen)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 804 SD = 102

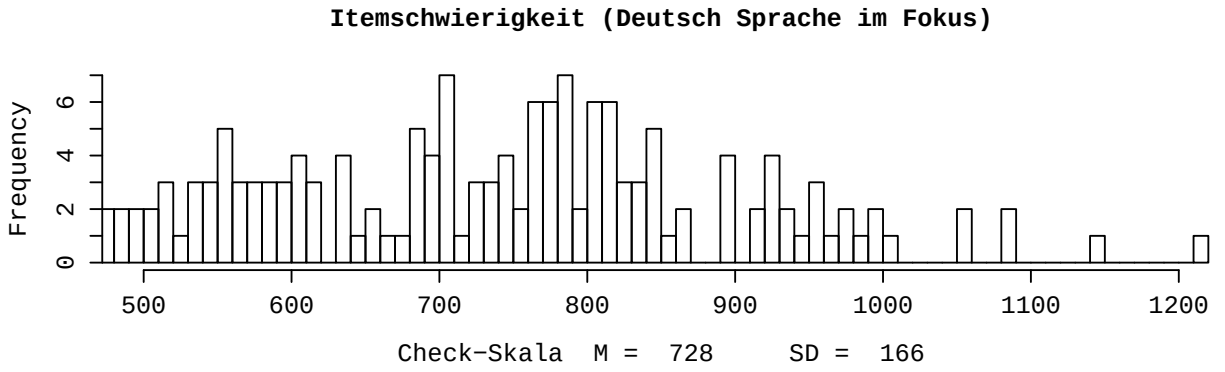

Itemschwierigkeit (Deutsch Lesen)
Frequency
0
2
4
6
8
12
500
600
700
800
900
1000
1100
1200
Check−Skala M = 762 SD = 141

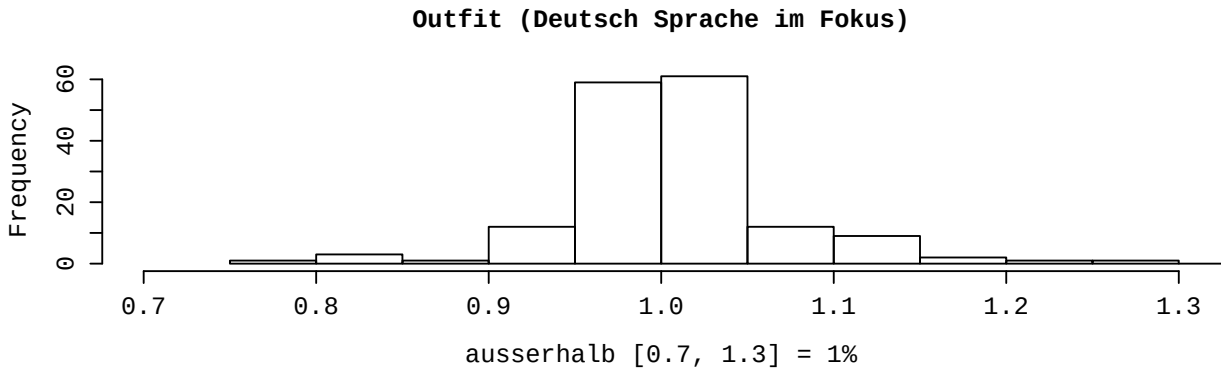

Outfit (Deutsch Lesen)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 4%

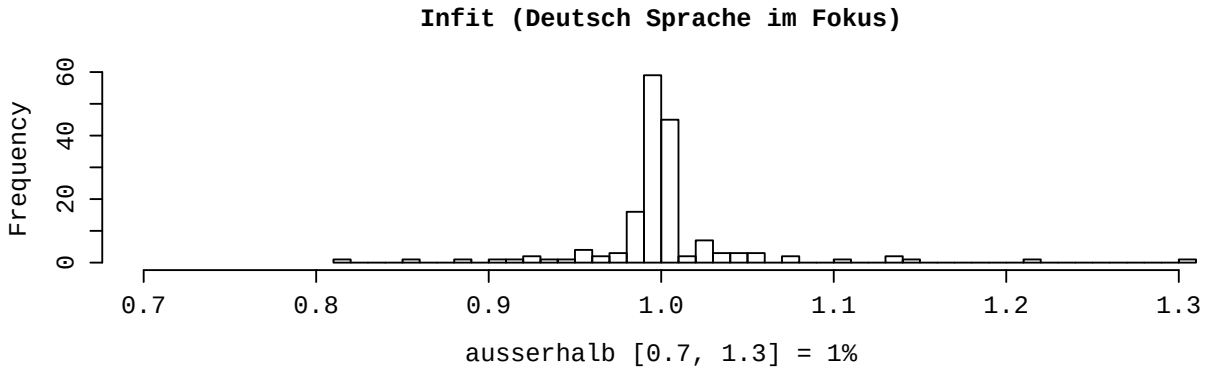

Infit (Deutsch Lesen)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

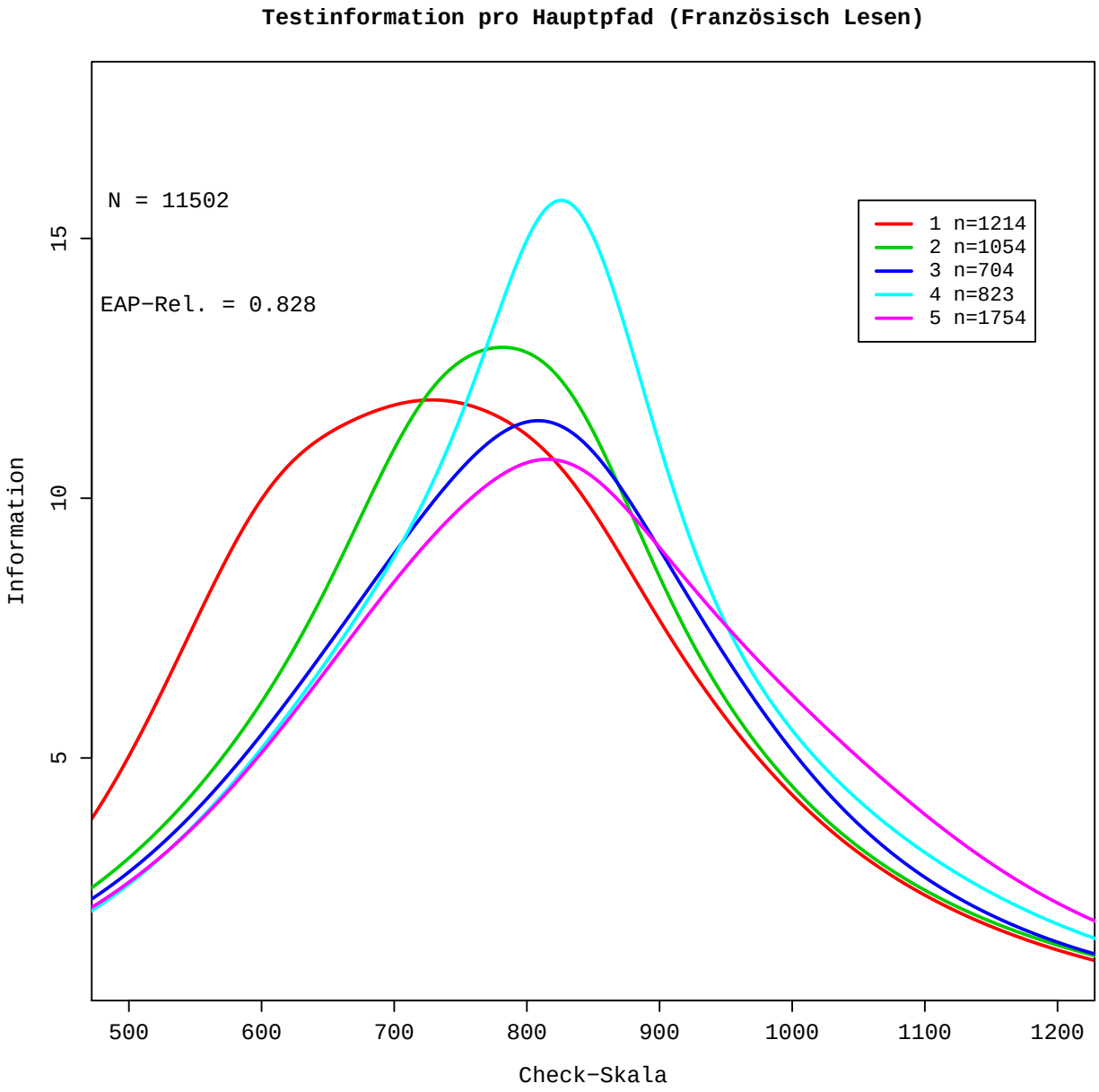

Testinformation pro Hauptpfad (Deutsch Sprache im Fokus)
N = 12790
EAP−Rel. = 0.865
1 n=1688
2 n=646
3 n=631
4 n=391
5 n=2385
Information
5
10
15
500
600
700
800
900
1000
1100
1200
Check−Skala

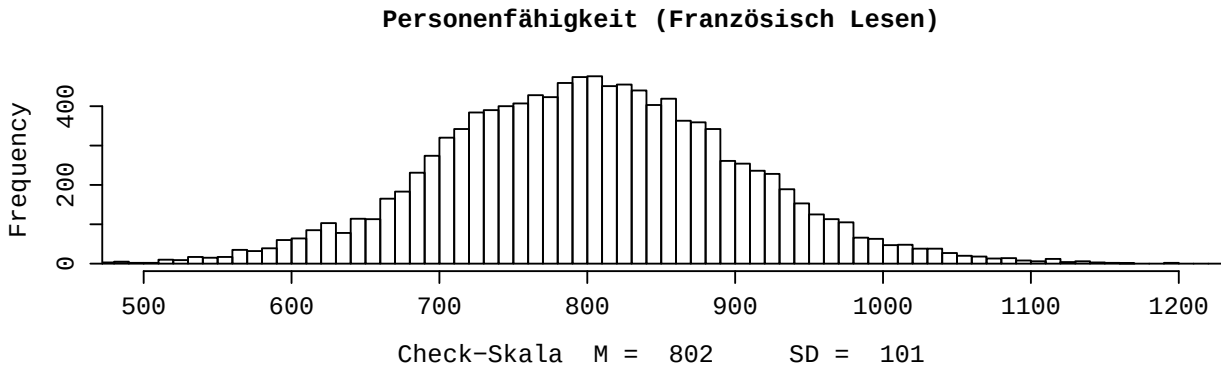

Personenfähigkeit (Deutsch Sprache im Fokus)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 803 SD = 123

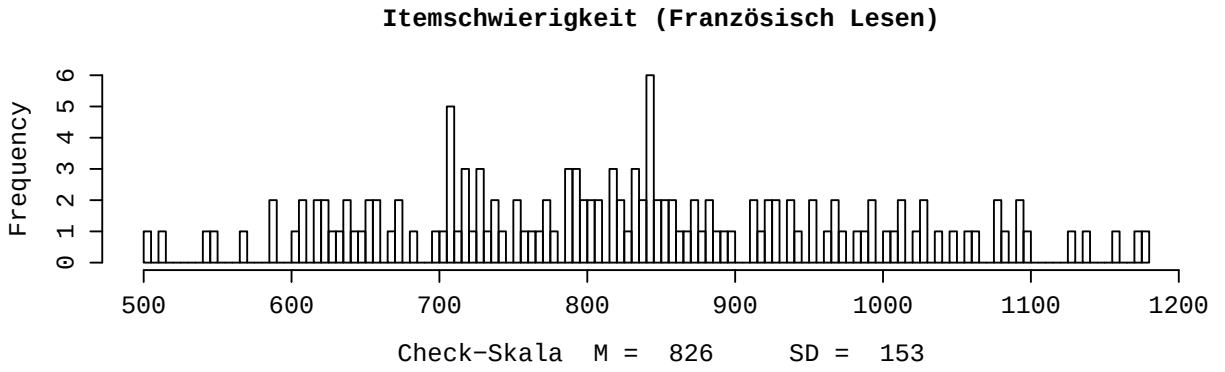

Itemschwierigkeit (Deutsch Sprache im Fokus)
Frequency
0
2
4
6
500
600
700
800
900
1000
1100
1200
Check−Skala M = 728 SD = 166

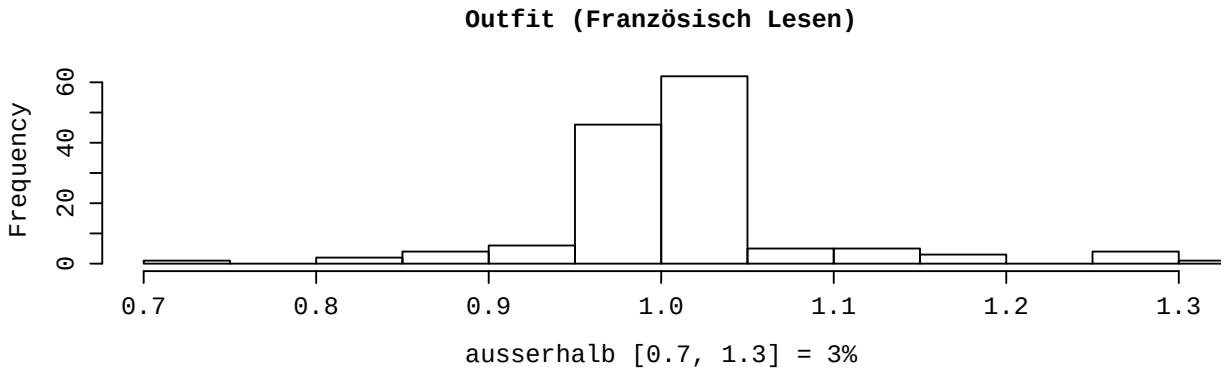

Outfit (Deutsch Sprache im Fokus)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

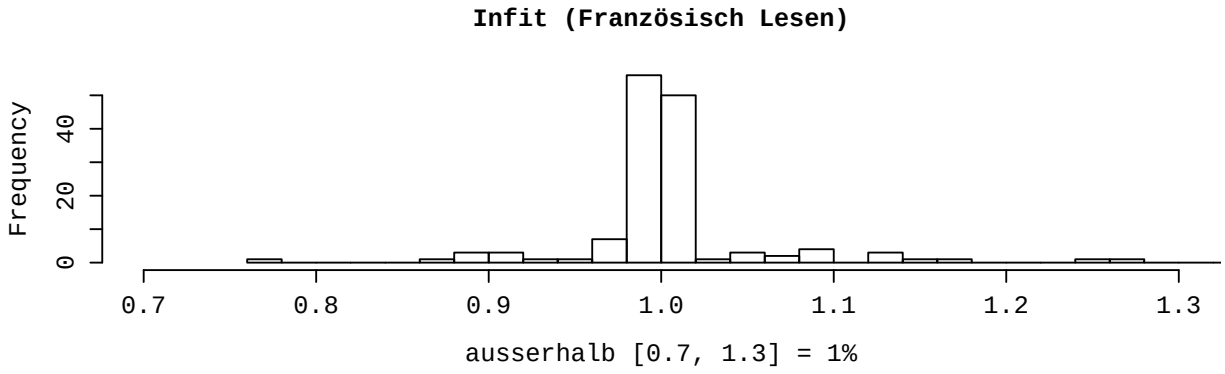

Infit (Deutsch Sprache im Fokus)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

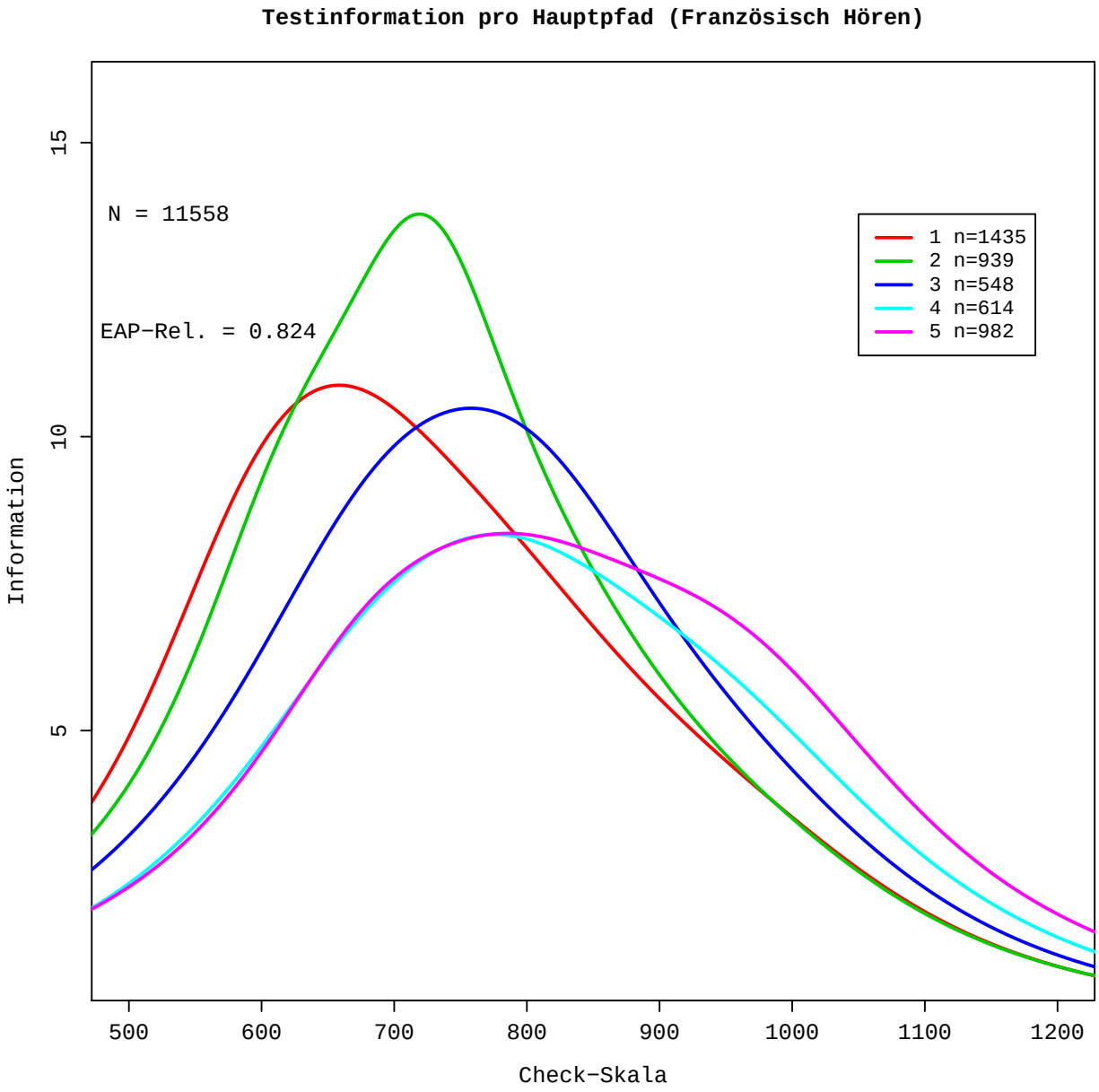

Testinformation pro Hauptpfad (Französisch Lesen)
N = 11502
EAP−Rel. = 0.828
1 n=1214
2 n=1054
3 n=704
4 n=823
5 n=1754
Information
5
10
15
500
600
700
800
900
1000
1100
1200
Check−Skala

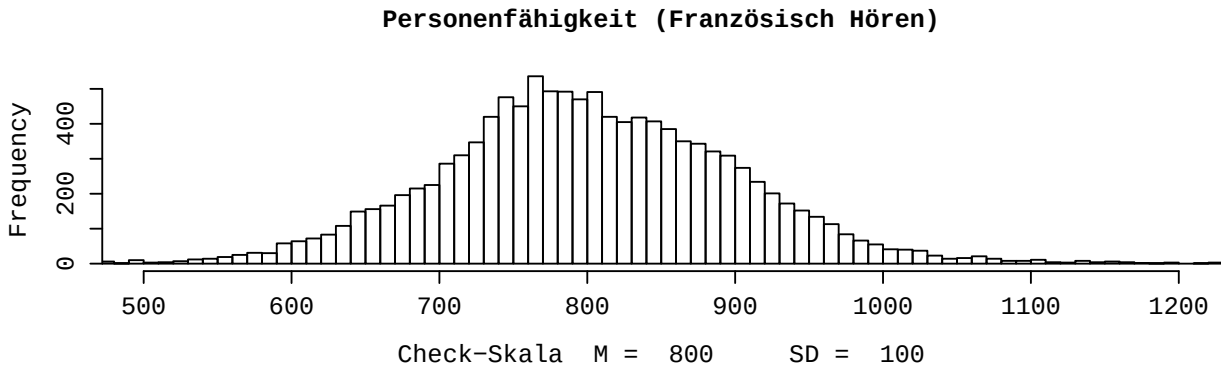

Personenfähigkeit (Französisch Lesen)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 802 SD = 101

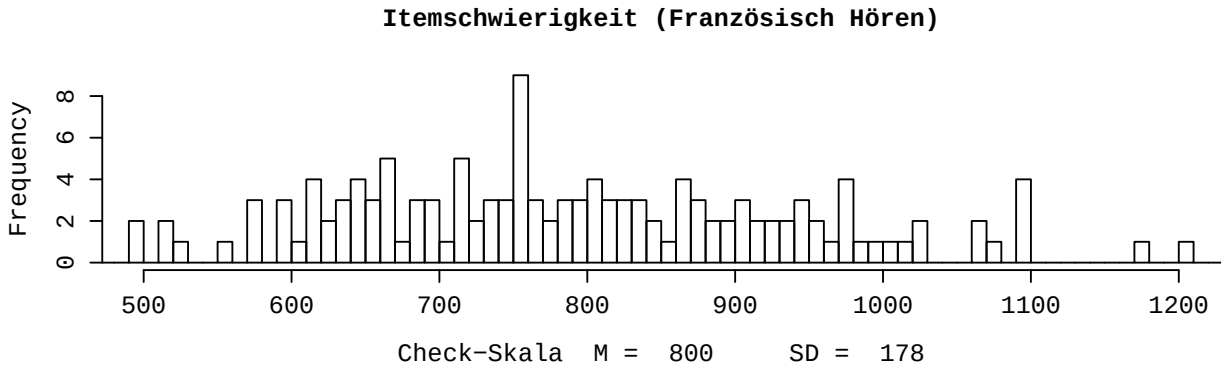

Itemschwierigkeit (Französisch Lesen)
Frequency
0
1
2
3
4
5
6
500
600
700
800
900
1000
1100
1200
Check−Skala M = 826 SD = 153

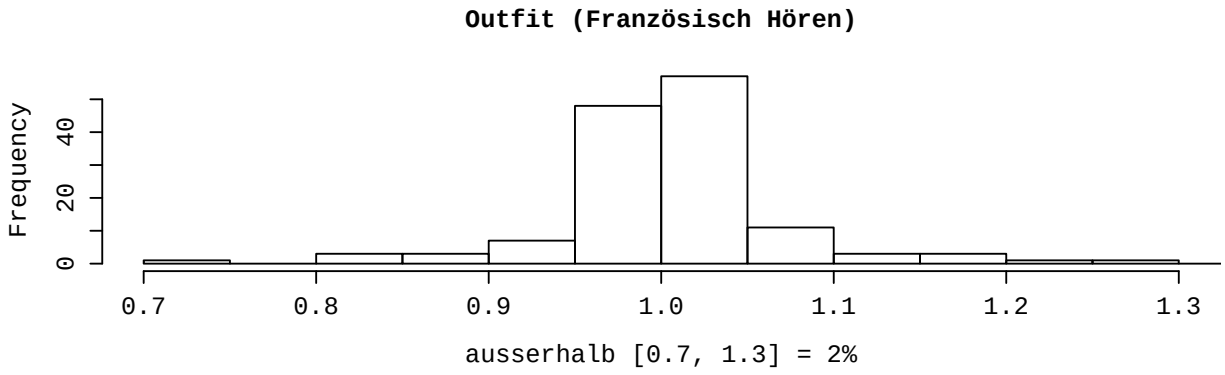

Outfit (Französisch Lesen)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 3%

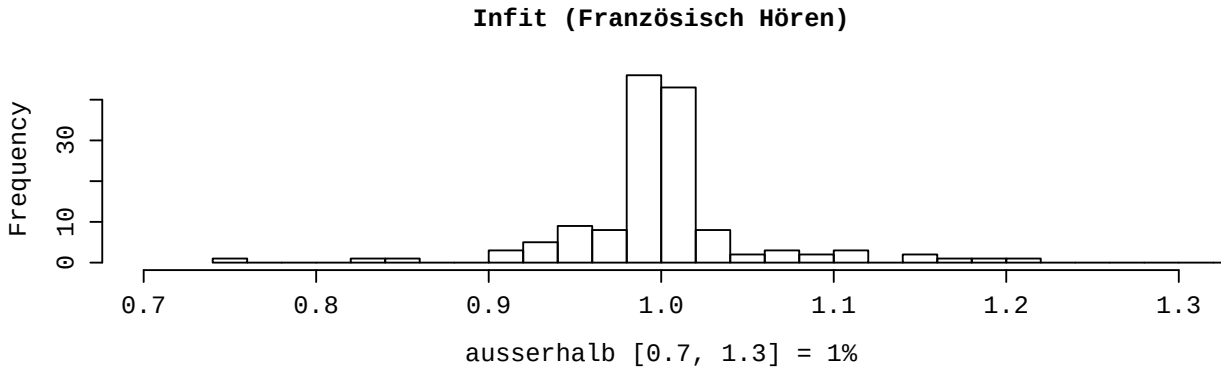

Infit (Französisch Lesen)
Frequency
0
20
40
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

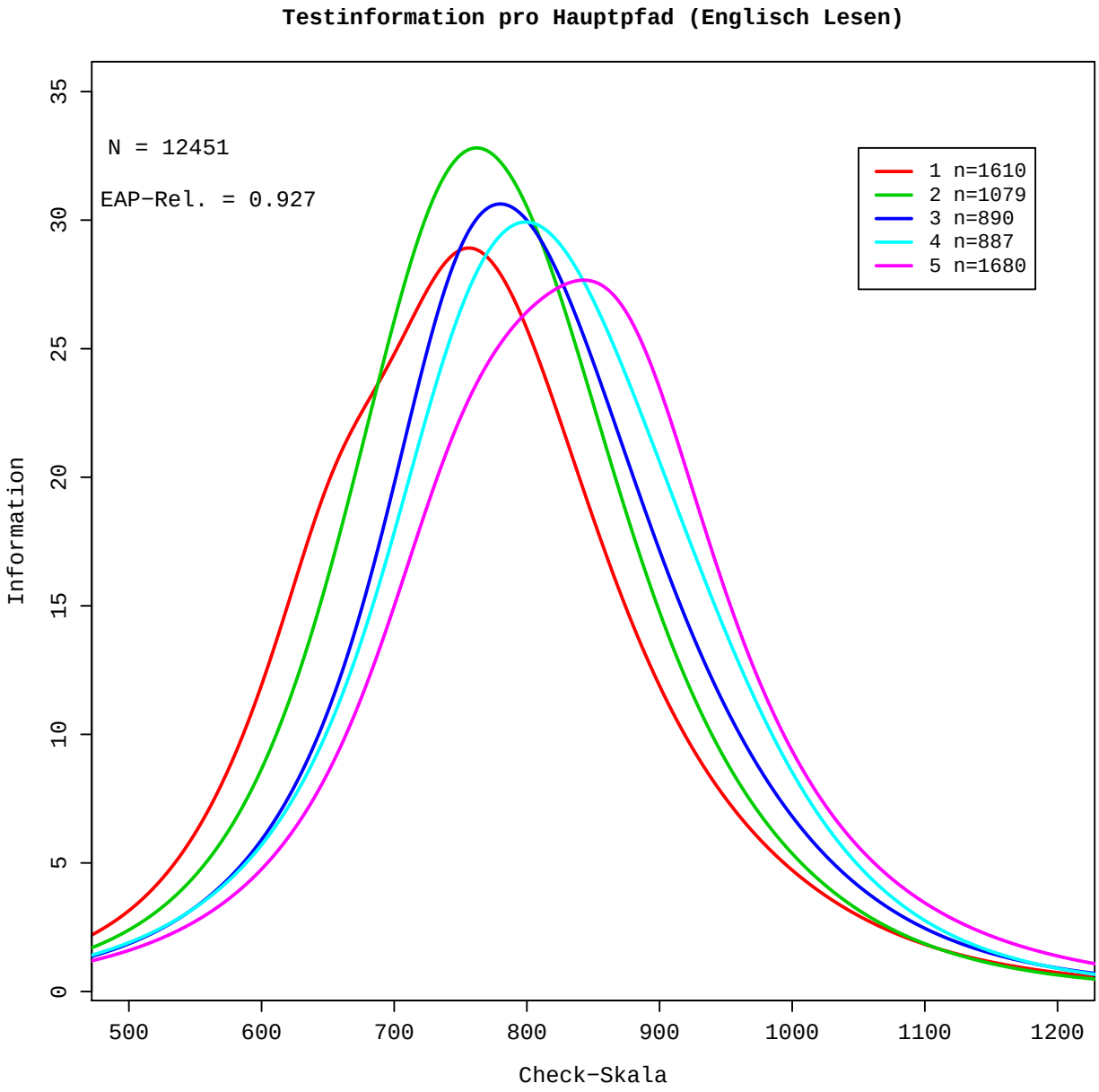

Testinformation pro Hauptpfad (Französisch Hören)
N = 11558
EAP−Rel. = 0.824
1 n=1435
2 n=939
3 n=548
4 n=614
5 n=982
Information
5
10
15
500
600
700
800
900
1000
1100
1200
Check−Skala

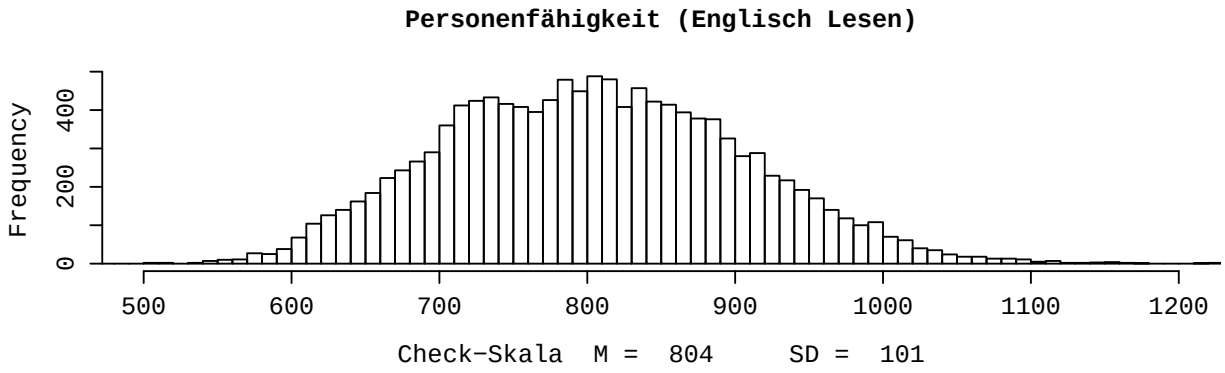

Personenfähigkeit (Französisch Hören)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 800 SD = 100

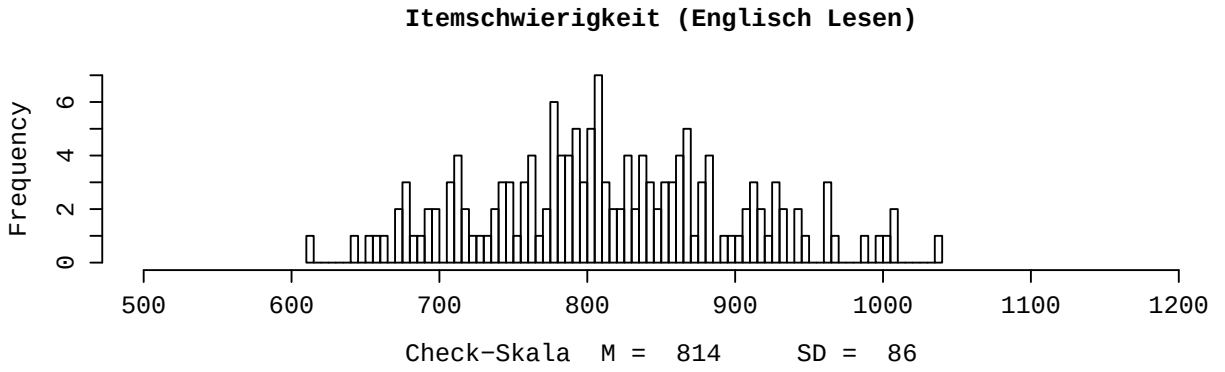

Itemschwierigkeit (Französisch Hören)
Frequency
0
2
4
6
8
500
600
700
800
900
1000
1100
1200
Check−Skala M = 800 SD = 178

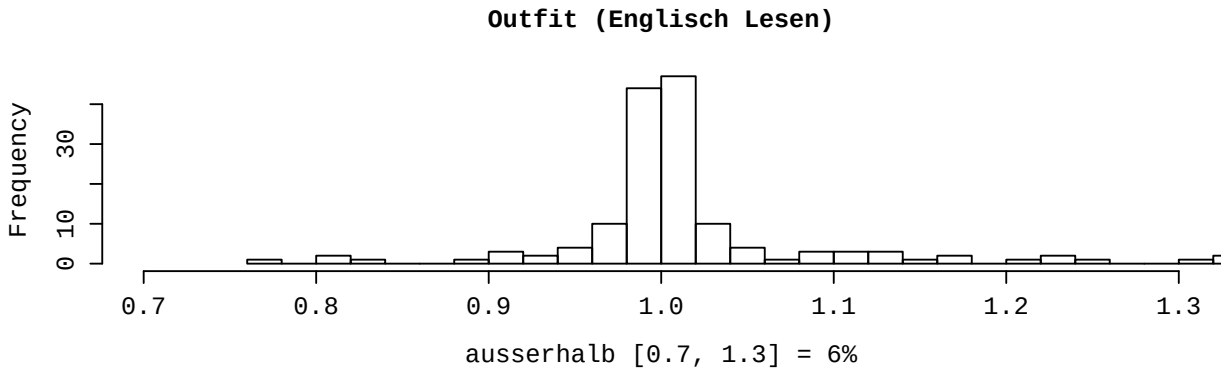

Outfit (Französisch Hören)
Frequency
0
20
40
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 2%

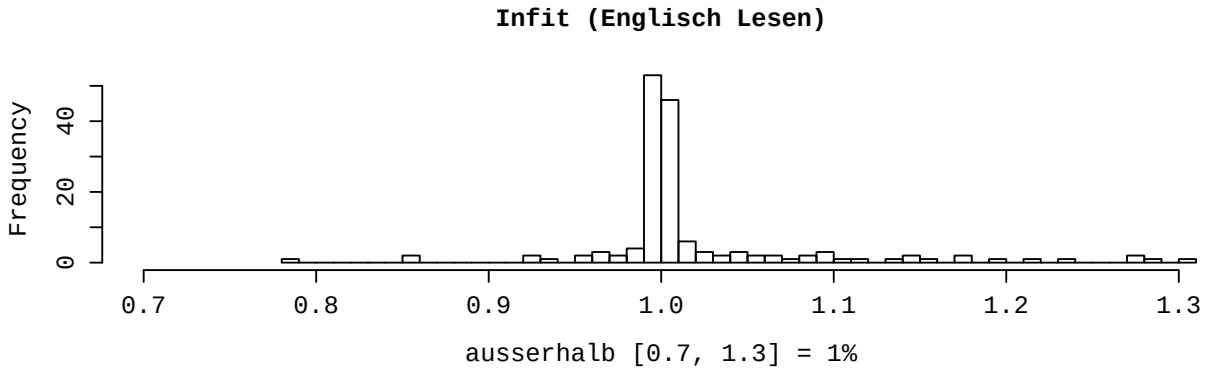

Infit (Französisch Hören)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

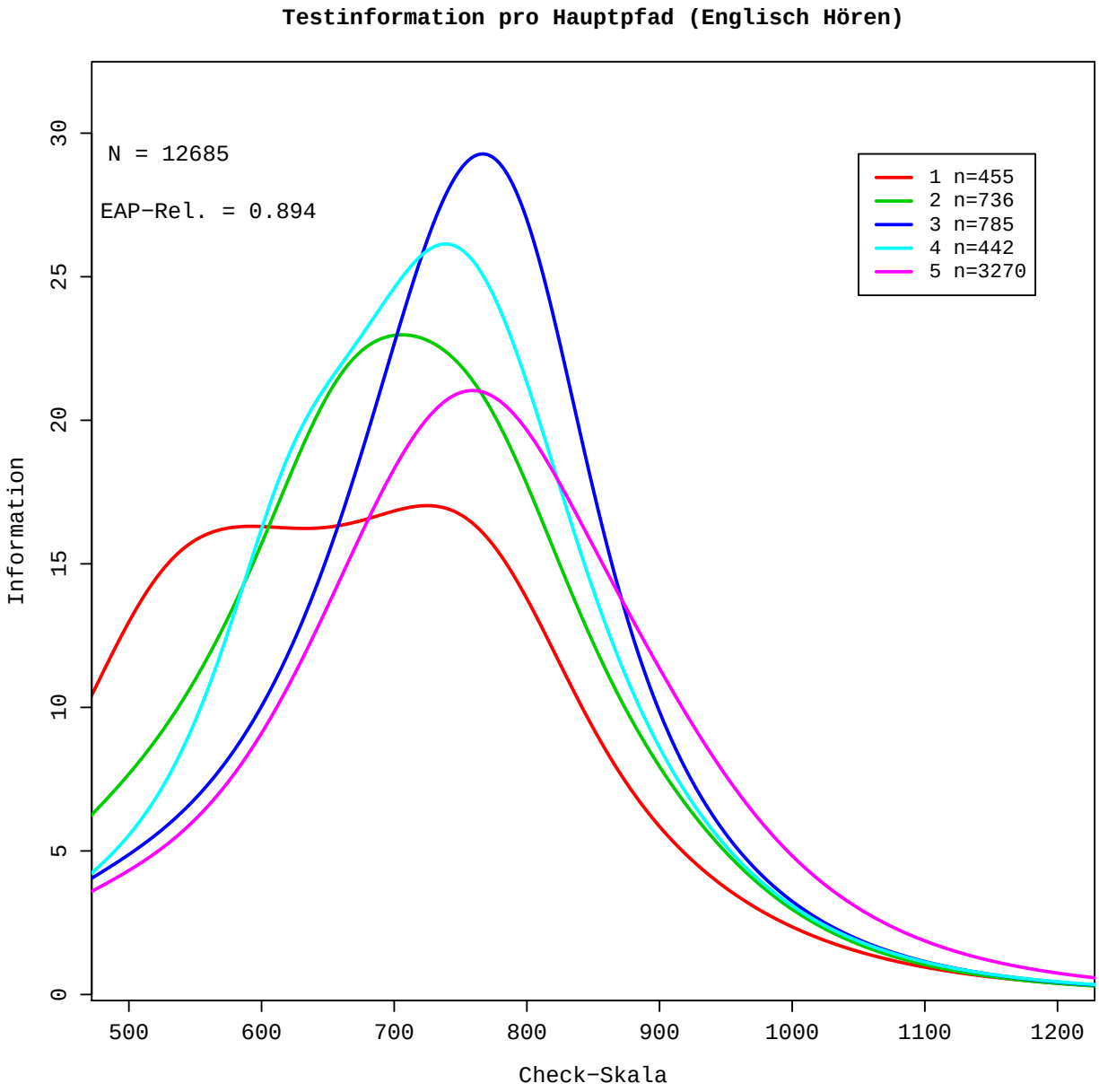

Testinformation pro Hauptpfad (Englisch Lesen)
N = 12451
EAP−Rel. = 0.927
1 n=1610
2 n=1079
3 n=890
4 n=887
5 n=1680
Information
0
5
10
15
20
25
30
35
500
600
700
800
900
1000
1100
1200
Check−Skala

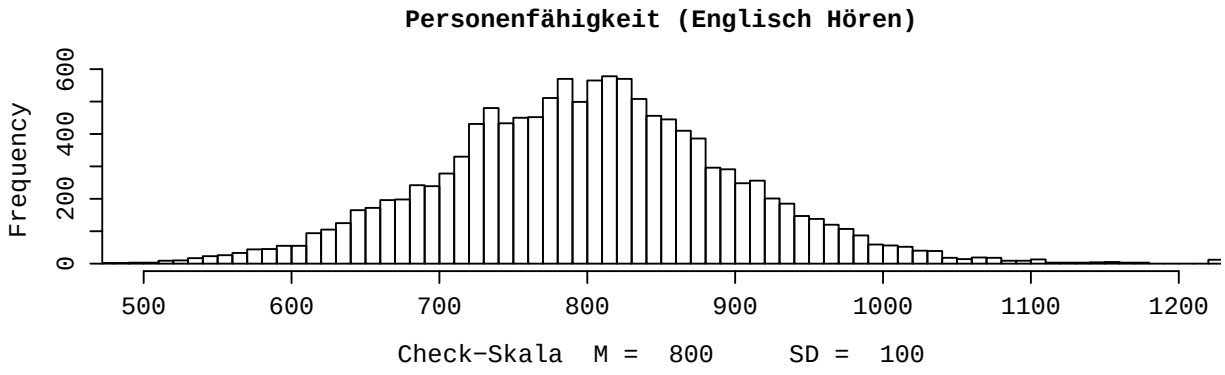

Personenfähigkeit (Englisch Lesen)
Frequency
0
200
400
500
600
700
800
900
1000
1100
1200
Check−Skala M = 804 SD = 101

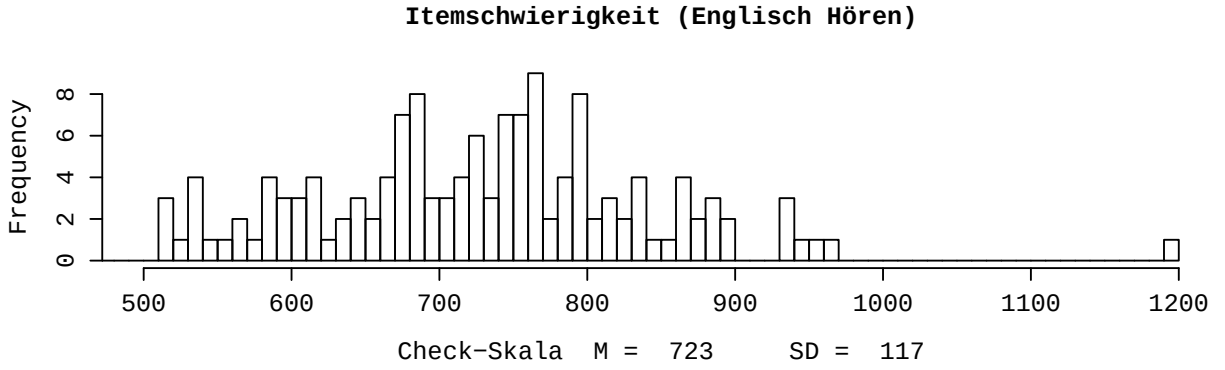

Itemschwierigkeit (Englisch Lesen)
Frequency
0
2
4
6
500
600
700
800
900
1000
1100
1200
Check−Skala M = 814 SD = 86

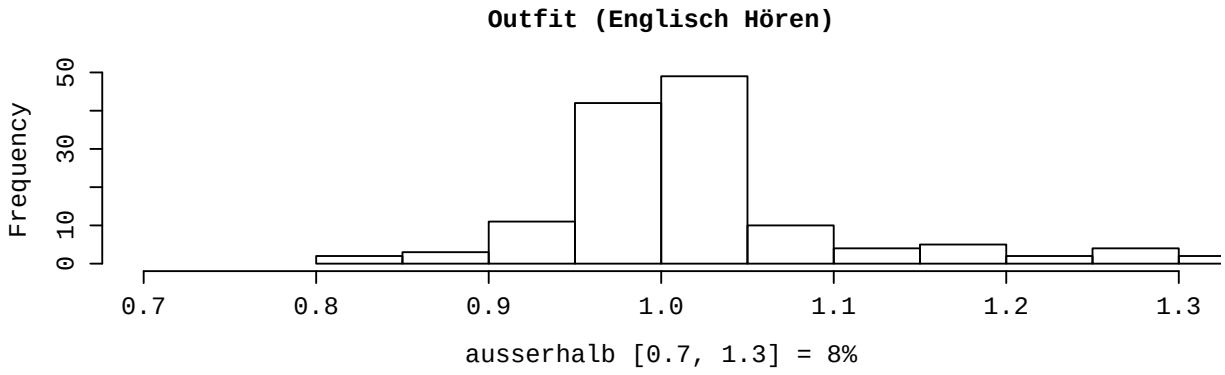

Outfit (Englisch Lesen)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 6%

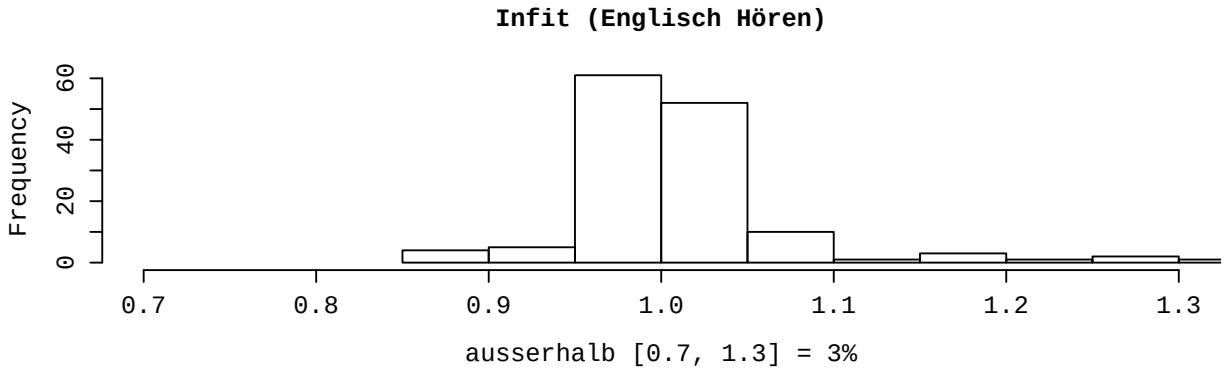

Infit (Englisch Lesen)
Frequency
0
20
40
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

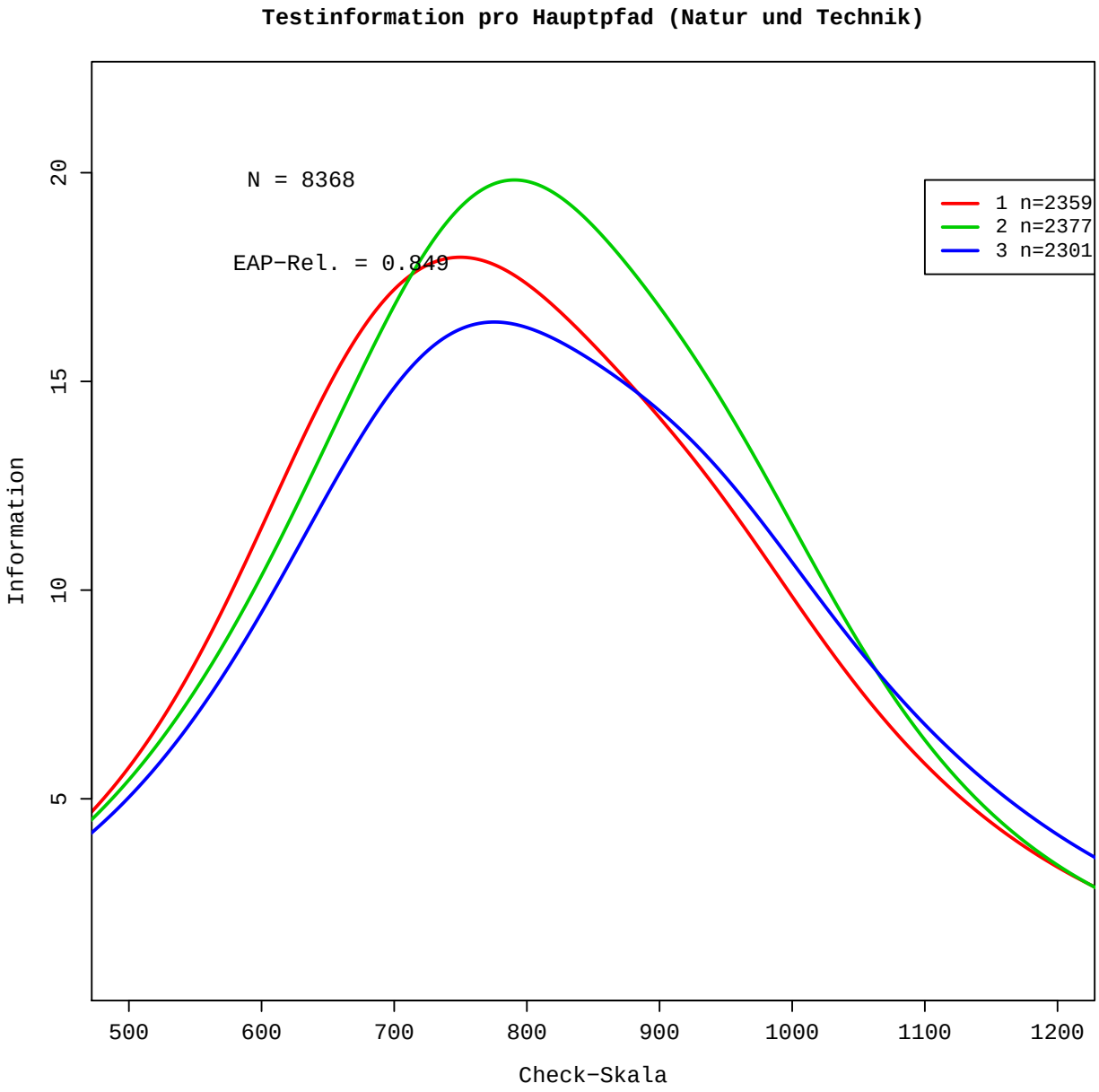

Testinformation pro Hauptpfad (Englisch Hören)
N = 12685
EAP−Rel. = 0.894
1 n=455
2 n=736
3 n=785
4 n=442
5 n=3270
Information
0
5
10
15
20
25
30
500
600
700
800
900
1000
1100
1200
Check−Skala

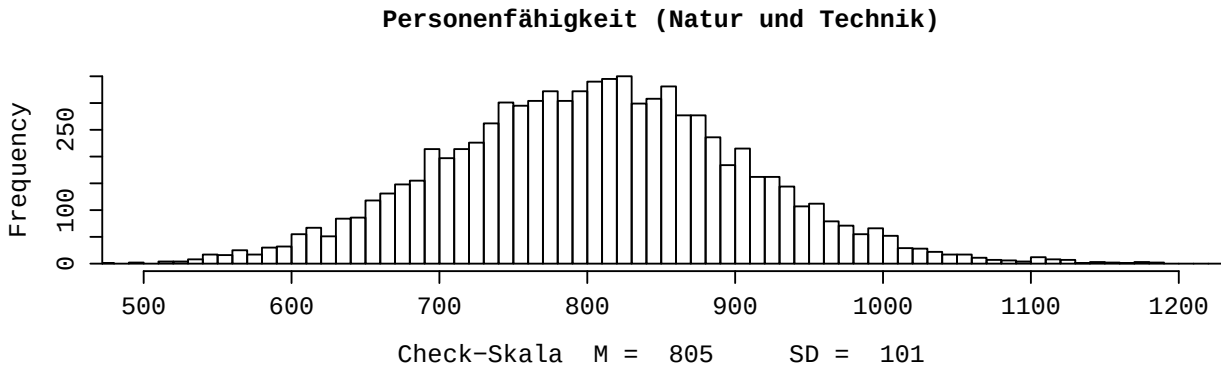

Personenfähigkeit (Englisch Hören)
Frequency
0
200
400
600
500
600
700
800
900
1000
1100
1200
Check−Skala M = 800 SD = 100

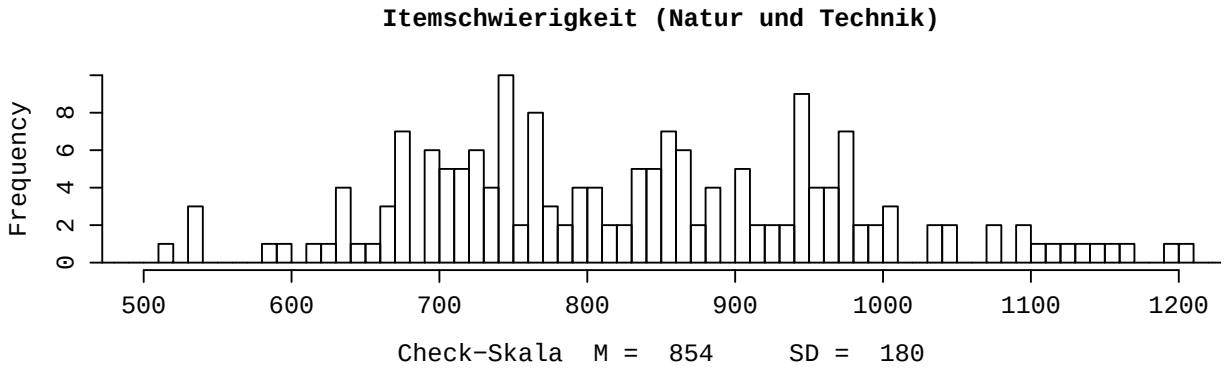

Itemschwierigkeit (Englisch Hören)
Frequency
0
2
4
6
8
500
600
700
800
900
1000
1100
1200
Check−Skala M = 723 SD = 117

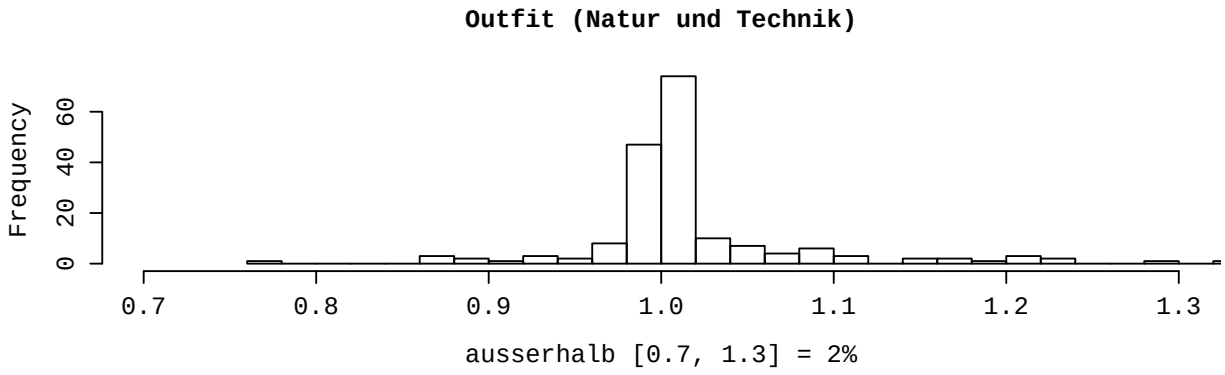

Outfit (Englisch Hören)
Frequency
0
10
30
50
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 8%


Infit (Englisch Hören)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 3%


Testinformation pro Hauptpfad (Natur und Technik)
N = 8368
EAP−Rel. = 0.849
1 n=2359
2 n=2377
3 n=2301
Information
5
10
15
20
500
600
700
800
900
1000
1100
1200
Check−Skala


Personenfähigkeit (Natur und Technik)
Frequency
0
100
250
500
600
700
800
900
1000
1100
1200
Check−Skala M = 805 SD = 101


Itemschwierigkeit (Natur und Technik)
Frequency
0
2
4
6
8
500
600
700
800
900
1000
1100
1200
Check−Skala M = 854 SD = 180


Outfit (Natur und Technik)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 2%

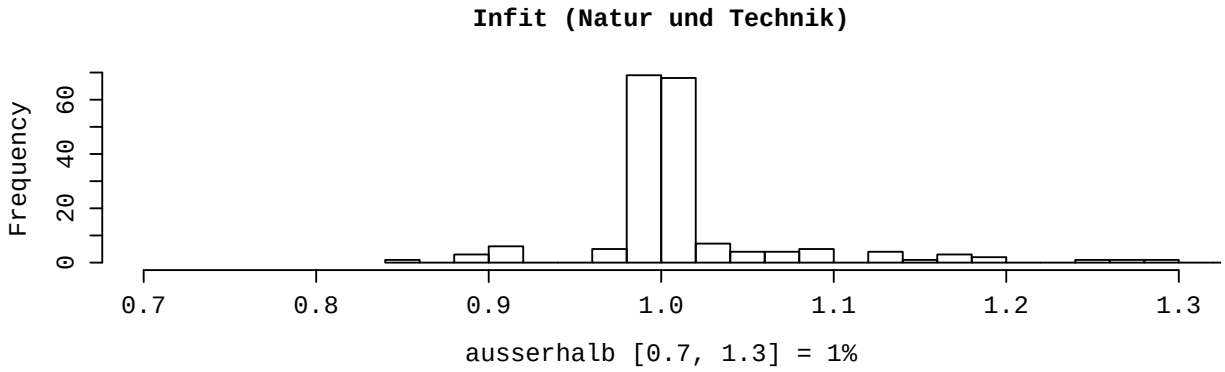

Infit (Natur und Technik)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

Tabelle 6: Mittelwerte und Standardabweichungen auf der Check-Skala im Check S2 2019

Skala	Mittelwert	Standardabweichung
Mathematik		
Zahl und Variable	801	119
Form und Raum	801	120
Grössen, Funktionen, Daten und Zufall	802	128
Deutsch		
Lesen	800	102
Sprache im Fokus	800	124
Schreiben	801	98
Französisch		
Lesen	800	99
Hören	800	99
Englisch		
Lesen	800	100
Hören	800	99
Schreiben	801	98
Natur und Technik	800	99

Tabelle 7: Skalen-Interkorrelationen und Fallzahlen im Check S2 2019

	mzuv	*mfur*	*mgfd*	*dles*	*dsif*	*dsch*	*fles*	*fhoe*	*eles*	*ehoe*	*esch*	*natw*
mzuv	12'856	0.76	0.79	0.6	0.66	0.51	0.55	0.47	0.55	0.5	0.42	0.67
mfur	12'856	12'856	0.79	0.59	0.64	0.51	0.54	0.45	0.52	0.48	0.4	0.67
mgfd	12'856	12'856	12'856	0.64	0.65	0.52	0.57	0.48	0.56	0.53	0.42	0.71
dles	12'747	12'747	12'747	12'824	0.67	0.62	0.65	0.54	0.68	0.62	0.5	0.73
dsif	12'702	12'702	12'702	12'747	12'777	0.65	0.62	0.53	0.61	0.55	0.54	0.65
dsch	12'722	12'722	12'722	12'737	12'695	12'841	0.5	0.46	0.49	0.45	0.47	0.56
fles	11'484	11'484	11'484	11'466	11'434	11'466	11'523	0.65	0.66	0.59	0.52	0.63
fhoe	11'492	11'492	11'492	11'472	11'441	11'475	11'516	11'532	0.55	0.52	0.46	0.52
eles	12'576	12'576	12'576	12'546	12'504	12'526	11'406	11'412	12'639	0.82	0.71	0.62
ehoe	12'581	12'581	12'581	12'551	12'509	12'533	11'406	11'413	12'628	12'645	0.71	0.57
esch	12'536	12'536	12'536	12'516	12'477	12'578	11'396	11'404	12'521	12'526	12'637	0.45
natw	8'313	8'313	8'313	8'294	8'277	8'299	7'892	7'898	8'246	8'247	8'226	8'352

Oberhalb Diagonale: Korrelationen nach Pearson (r)
Unterhalb Diagonale: Anzahl Fälle
mzuv = Mathematik Zahl und Variable, *mfur* = Mathematik Form und Raum, *mgfd* = Mathematik Grössen, Funktionen, Daten und Zufall, *dles* = Deutsch Lesen, *dsif* = Deutsch Sprache im Fokus, *dsch* = Deutsch Schreiben, *fles* = Französisch Lesen, *fhoe* = Französisch Hören, *eles* = Englisch Lesen, *ehoe* = Englisch Hören, *esch* = Englisch Schreiben, *natw* = Natur und Technik