Technischer Bericht zum Check S3 2018

Stéphanie Berger, Laura Helbling, Martin J. Tomasik, Urs Moser

Institut für Bildungsevaluation (IBE)
Assoziiertes Institut der Universität Zürich
Zürich, 19. Juni 2018

1. Einleitung

Die Kantone Aargau, Basel-Stadt, Basel-Landschaft und Solothurn beauftragten das Institut für Bildungsevaluation, Assoziiertes Institut der Universität Zürich, mit der Entwicklung und Durchführung gemeinsamer Leistungstests in der 3. und 6. Klasse der Primarstufe (Check P3 und Check P6) sowie in der 2. und 3. Klasse der Sekundarstufe I (Check S2 und Check S3). 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 S3 2018.

2. Testdesign

Der Check S3 2018 umfasste insgesamt zwölf Tests, wobei drei verschiedene Testformate unterschieden werden: Online-Tests, Papier-Tests und mündliche Tests. Tabelle 1 enthält eine Übersicht über die erfassten Kompetenzbereiche, die Testformate sowie die Anzahl Aufgaben für die Online-Tests pro Test oder Testteil.

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
Schreiben	Papier	---	---	---	---	---
Sprechen	Mündl.	---	---	---	---	---
Englisch						
Lesen	Online	7	7	8	8	30
Hören	Online	7	7	8	8	30
Natur und Technik[a]						
Allgemeine Naturwissenschaften	Online	20	---	---	---	20
Physik, Chemie oder Biologie	Online	5	5	---	---	10

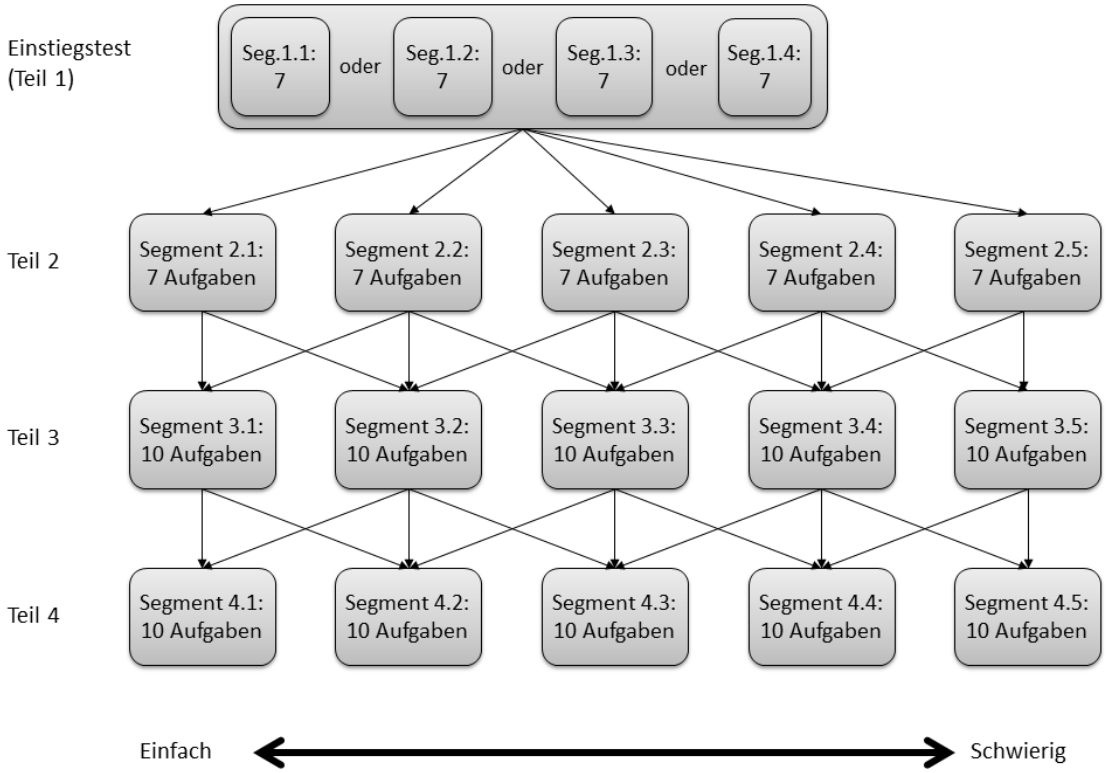
a) Der Test in Natur und Technik besteht aus einem allgemeinen Teil mit 20 Aufgaben sowie zwei weiteren Zwei-Stage-Tests in Physik, Chemie oder Biologie mit je 10 Aufgaben.

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* sowie *Natur und Technik*. 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 vor 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). Im Hintergrund 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 (abgesehen von *Natur und Technik*) äquivalent für alle Skalen, die online getestet werden. Lediglich die Aufgabenzahl pro Testteil unterscheidet sich zwischen diesen Online-Tests. Die Online-Tests in *Natur und Technik*, die von diesem Testdesign abweichen, werden nachfolgend in einem eigenen Abschnitt beschrieben.

Abbildung 1. Multistage-Design des Online-Tests Deutsch Sprache im Fokus des Check S3



Jeder Multistage-Test umfasste vier Testteile (Teil 1-4, Abbildung 1) und begann mit dem ersten Testteil (Einstiegstest), dessen Aufgaben von eher einfacherer Schwierigkeit waren. Der erste Testteil bestand aus vier äquivalenten Testsegmenten und die Schülerinnen und Schüler wurden zufällig einem der vier äquivalenten Testsegmenten zugewiesen. Für die Testteile 2 bis 4 standen jeweils fünf Segmente von 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[1] (Duanli et al., 2014). Die Aufgaben eines Testteils mussten innerhalb einer vorgegebenen Zeit gelöst werden. Je nach Testteil und Skala variiert diese Zeit zwischen fünf 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 Lösung zu korrigieren. Nach Ablauf der Zeit konnten die Aufgaben des Testteils nicht mehr weiterbearbeitet werden.

[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 z.B. entspricht einem Pfad, bei dem Schülerinnen und Schüler Testteil 2 bis 4 auf der leichtesten Schwierigkeitsstufe lösen. Der Hauptpfad 5 entspricht demgegenüber einem Pfad, bei dem Testteile 2 bis 4 auf der höchsten Schwierigkeitsstufe gelöst wurden.

Praktische Arbeit und Online-Test in Natur und Technik

Der Check in Natur und Technik bestand aus einer praktischen Arbeit, die im Klassenverband durchgeführt wird, und drei anschliessenden Online-Tests. Der erste Online-Test enthielt Aufgaben, die sich auf die vorab durchgeführte praktische Arbeit beziehen, und beinhaltet zudem Aufgaben zum Thema „Allgemeine Naturwissenschaften“. Im Check S3 2018 bezog sich die praktische Arbeit auf Einflussfaktoren der Schwingungsdauer eines Pendels. Für den zweiten und dritten Online-Test standen sechs Themen aus drei verschiedenen Fachbereichen 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 je nachdem, wie gut sie diese lösten, wurden ihnen anschliessend fünf eher einfache oder fünf eher schwierige Aufgaben zugewiesen.

Papier-Tests (Deutsch Schreiben und Französisch Schreiben)

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, Brief, 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 Versionen eingesetzt wird, entscheidet die Lehrperson. Beide Versionen umfassen zwei Aufgaben (Schreibaufträge), die beide von den Schülerinnen und Schülern gelöst werden müssen. Wie in Deutsch verlangen die Aufgabenstellungen jeweils verschiedene Textsorten (u.a. Brief, Mitteilung, Erzählung, Reportage). Im Check S3 2018 wurden Schreibkompetenzen in Deutsch und Französisch getestet.

Mündliche Tests (Englisch und Französisch)

In den Fächern Englisch und Französisch wurden zusätzlich mündliche Kompetenzen getestet. Die Teilnahme an diesen mündlichen Tests war freiwillig. Es wurden Einzelgespräche mit den Schülerinnen und Schülern geführt. Den Lehrpersonen wurden hierfür drei Aufgabenbeispiele zur Verfügung gestellt, die eine i) (nicht bewertete) Einstiegsaufgabe, ii) eine Aufgabe zum monologischen Sprechen und iii) eine Aufgabe zum dialogischen Sprechen umfassen.

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 S3 2018 teilgenommen hatten. Tabelle 2 beinhaltet ebenfalls die Anzahl rückgemeldeter Online- und Papier-Tests pro Kanton.

Tabelle 2. Teilnahme

Kantone	*Schulen*	*Klassen*	*Schülerinnen und Schüler*	*Anzahl Online-Tests*	*Anzahl Papier-Tests*
Aargau	109	393	6’509	42’828	11’301
Basel-Landschaft	21	150	2’612	20’332	5’087
Solothurn	29	112	1’867	14’444	3’635
Total	159	655	10’988	77’604	20’023

5. Testdurchführung

Testzeitfenster

Die Durchführung des Checks S3 fand zwischen dem 16. April und dem 25. Mai 2018 statt. Die Ergebnisse wurden am 15. Juni 2018 zurückgemeldet.

Anzahl durchgeführter Tests

Während des gesamten Zeitfensters nahmen 10’988 Schülerinnen und Schüler aus den Kantonen Aargau, Basel-Landschaft und Solothurn am Check S3 2018 teil Die Schülerinnen und Schüler lösten in den Fächern Deutsch, Englisch, Französisch, Mathematik sowie Natur und Technik insgesamt 77’604 Online-Tests [2]. Zusätzlich wurden 10’887 Texte in Deutsch und 9’136 Texte in Französisch verfasst und beurteilt. Zudem beurteilten die Lehrpersonen die Sprechkompetenzen von 124 Schülerinnen und Schülern in Englisch und 91 Schülerinnen und Schülern in Französisch.

6. Online-Tests: Auswertung

Scoring

Das Scoring der Online-Tests, das heisst, das Übersetzen der Antworten in numerische Werte, wurde direkt im Onlinesystem (assessment delivery platform) vorgenommen. Dabei wurde zwischen richtig, falsch und nicht bearbeiteten Aufgaben unterschieden. Das ordnungsgemässe 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. Das vom System vorgenommene Scoring der offenen Aufgaben wurde im Rahmen der

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

Auswertung manuell nachgeprüft, da zum Teil nicht alle möglichen richtigen Antwortalternativen im System erfasst sind und diese daher laufend ergänzt werden müssen. Vor der Berechnung der Ergebnisse der Online-Tests wurde die Funktionalität 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 bearbeitet und mindestens 20 % aller Aufgaben gelöst hatte. Wurden alle Aufgaben eines Testteils nicht bearbeitet, dann wurde der entsprechende Testteil für die Berechnung der Testergebnisse nicht einbezogen.
Neben der Punktzahl für jeden Kompetenzbereich wurden fachübergreifende Gesamtwerte ausgewiesen. Die Gesamtwerte in Deutsch und Mathematik wurden als arithmetische Mittelwerte aus den Ergebnissen in den einzelnen Kompetenzbereichen berechnet. Die Gesamtwerte in Englisch und Französisch wurden als arithmetische Mittelwerte aus den Ergebnissen in den einzelnen Kompetenzbereichen, aber ohne Ergebnis im Sprechen, berechnet. Generell wurden die Gesamtwerte nur dann berechnet, wenn für die jeweiligen Kompetenzbereiche allesamt gültige Testergebnisse vorlagen, die miteinander verrechnet werden konnten.

Skalierung und Parameterschätzung

Die Skalierung der Daten aus den Online-Tests erfolgte mit dem Softwarepaket TAM von Kiefer, Robitzsch und Wu (2017) in der Entwicklungsumgebung R (R Development Core Team, 2016). Für die Auswertung der dichotomen Aufgaben beziehungsweise Items kam ein Zwei-Parameter-Logistisches-Modell (auch: Birnbaum-Modell; 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

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

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 (vgl. Rasch-Modell; Rasch, 1960), bei dem die Steigungsparameter über alle Items gleichgesetzt werden, ergibt sich 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 stärker gewichtet als wenig trennscharfe Items für die Fähigkeitsschätzung (vgl. Birnbaum, 1968).

Die Kalibrierung der Aufgaben wurde basierend auf Tests von Schülerinnen und Schülern durchgeführt, die diese zu einem vorher definierten Stichdatum (meist kurz vor dem Ende des Testzeitfensters) durchgeführt hatten. Es wurde darauf geachtet, dass der Kalibrierungsdatensatz den grössten Teil (in der Regel >90%) der Tests der teilnehmenden Schülerinnen und Schüler beinhaltete.

In diesem Kalibrierungsdatensatz wurden alle Schülerinnen und Schüler ausgeschlossen, die mehr als 20% fehlende Werte auf den Items des jeweiligen Tests hatten. Dieser Ausschluss betraf nur einen kleinen Teil der Schülerinnen und Schüler (in der Regel < 5%) und soll die Stabilität der Parameterschätzungen erhöhen.

Die Kalibrierung erfolgte in einem bis mehreren Schritten, in denen in der Regel Items mit geringer oder sogar negativer CTT-Trennschärfe von r_{tt} <.20, berechnet als punkt-biseriale Korrelation zwischen dem Einzelitem und dem Gesamttest, und solche 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 weiter 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 in Zusammenhang mit ii) inhaltlichen Überlegungen. Pro Test mussten nur wenige Aufgaben ausgeschlossen werden, wie Tabelle 3 zeigt.

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 nach Jahr unterschiedlichen Tests jeweils ein Teil gleicher Aufgaben wiederholt eingesetzt wird. Die Itemparameter bereits verwendeter Aufgaben wurden dementsprechend nicht neu geschätzt, sondern zur Skalenstabilisierung fixiert, das heisst auf den bisher geschätzten Itemparameterwerten belassen (Link-Items). Dies ermöglichte es, die Schwierigkeit und die Trennschärfe der neuen Aufgaben auf der bestehenden Skala einzuordnen. Die mehrfach eingesetzten Aufgaben ermöglichten 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 Parameter können linear auf eine beliebige Skala, z.B. auf die Check-Skala von 200 bis 1200 Punkten, transformiert werden. Für den Check S3 wurde der Mittelwert der Schülerinnen und Schüler im ersten Erhebungsjahr – das ist das Jahr 2015 – auf 800 Punkte und die Standardabweichung auf 60 Punkte standardisiert. Die Ergebnisse des diesjährigen Checks können dank des Testlinkings mit den Ergebnissen der Vorjahre verglichen werden.

Zur Sicherstellung der Qualität des Testlinkings wurden die mehrfach verwendeten Testaufgaben (Link-Items) auf Differential Item Functioning (DIF) überprüft. Das heisst, es wurde überprüft, ob die mehrfach verwendeten Aufgaben in unterschiedlichen Schülerjahrgängen gleich funktionieren oder ob bei gleichen Schülerfähigkeiten nach Jahrgang unterschiedliche Lösungswahrscheinlichkeiten für diese Link-Items vorliegen. Beispielsweise könnte es vorkommen, dass in einem Schülerjahrgang über ein verstärktes Üben bestimmter Aufgabenformate/Inhalte gewisse Aufgabenstellungen von

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

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 der empirisch 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 aktuellen Check pro Skala als Link-Items mit fixierten Parametern verwendet wurden, sowie 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 und der Skalen-Interkorrelationen.

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

Skala	*Anzahl Items Total*	*Anzahl Ausschlüsse*	*Anzahl Link-Items*	*EAP-Reliabilität*
Mathematik	231	14	167	0.93
Deutsch				
Lesen	190	6	170	0.90
Sprache im Fokus	163	8	132	0.90
Französisch				
Lesen	143	7	125	0.88
Hören	143	9	120	0.86
Englisch				
Lesen	143	8	126	0.93
Hören	143	4	134	0.91
Natur und Technik	110	8	51	0.83

7. Papier-Tests: Auswertung Schreiben

Kriterien zur Beurteilung der Texte

Die Texte in *Deutsch Schreiben* und *Französisch 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 Französischtexte umfasst zwei Dimensionen:

1. Inhalt
2. Sprachrichtigkeit

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

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 diesem Vorgehen sollte ein einheitlicher Beurteilungsmassstab und damit eine Beurteilungskonsistenz erreicht werden. Dass dies gelungen ist, zeigen die Auswertungen der Doppelkorrekturen in Tabelle 4 für Deutsch und Tabelle 5 für Französisch. Bestimmt wurde die Intraklassenkorrelation r_{ICC} (genauer: „oneway multiple raters consistency“ nach McGraw und 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} immer über .90 ohne augenfällige Unterschiede zwischen den Fächern Deutsch und Französisch.

Tabelle 4. Beurteilerübereinstimmung Deutsch Schreiben

Skala	$N_{subjects}$	N_{rater}	r_{ICC}	$CI_{95\%}$	P(r > .90)
Inhalt	128	9	.942	[.925;.956]	<.001
Textaufbau	128	9	.937	[.920;.953]	<.001
Sprachrichtigkeit	128	9	.972	[.965;.979]	<.001
Sprachangemessenheit	127	9	.963	[.952;.972]	<.001

Tabelle 5. Beurteilerübereinstimmung Französisch Schreiben

Skala	$N_{subjects}$	N_{rater}	r_{ICC}	$CI_{95\%}$	P(r > .90)
Aufsatz 1					
Inhalt	102	8	.970	[.960;.978]	<.001
Sprachrichtigkeit	102	8	.975	[.968;.982]	<.001
Aufsatz 2					
Inhalt	102	8	.982	[.976;.987]	<.001
Sprachrichtigkeit	102	8	.975	[.967;.982]	<.001

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 relativ einfach als richtig oder falsch korrigiert werden können, spielt der Beurteilungsmassstab der beurteilenden Personen bei offenen Aufgabenformaten eine Rolle für das Testergebnis. Beurteilt beispielsweise Beurteiler/in A systematisch strenger als Beurteiler/in B, dann ist dies für all jene Schülerinnen und Schüler ungerecht, deren Text von Beurteiler/in A beurteilt wird. Wird die Strenge oder Milde in der Beurteilung der Texte bei der Berechnung der Testergebnisse nicht berücksichtigt, dann wird ein Text je nach Beurteiler/in entweder besser oder weniger gut beurteilt. Aus diesem Grund wurde die Beurteilungsstrenge der beurteilenden Personen (Rater) als Facette der Urteilssituation aufgefasst und bei der Berechnung der Ergebnisse wie folgt berücksichtigt[4]:

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

[4] Bei der Skalierung wurde hingegen 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.

P_{inm} und P_{inm-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* und β_j der Schwierigkeit des Items 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 OutfitStatistiken 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 Französisch 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 Französisch Schreiben wurde auf den Mittelwert der beiden Kompetenzbereiche Französisch Lesen und Französisch Hören standardisiert. Die Standardabweichung wurde für beide Papier-Tests auf 60 Punkte festgelegt. 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. Mündliche Tests: Auswertung Sprechen

Die Beurteilung der sprachlichen Kompetenzen der Schülerinnen und Schüler wurde von den Lehrpersonen selbst nach einem standardisierten Verfahren vorgenommen. Den Lehrpersonen wurden hierfür Audio-Dateien mit vier Vergleichsbeispielen pro Sprechaufgabe sowie ein Beurteilungsbogen zur Verfügung gestellt. Auf dem Beurteilungsbogen konnten die Lehrpersonen die Schülerinnen und Schüler in Abgleich mit den Sprechbeispielen auf einer neunstufigen Skala einordnen. Für die Ergebnisrückmeldung wurde aus den beiden Beurteilungen der monologischen und der dialogischen Aufgabenstellung das arithmetische Mittel der erreichten Punktzahl berechnet.

9. 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 rückgemeldet 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 % 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. Es ist somit etwas arbiträr, wer die Aufgabe korrekt löst.

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. Überprüfung 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 Itemschwierigkeiten, 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. D.h., neben der zu messenden Personenfähigkeit beeinflussen weitere Faktoren die Lösungswahrscheinlichkeiten der entsprechenden Aufgaben. Die Abwesenheit von DIF ist ein zu prüfendes Gütekriterium der Itemanalyse.

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 Item-Charakteristik-Kurve basierend auf den beobachteten Antworten und der gemäss Modell erwarteten Steigung der Item-Charakteristik-Kurve. Die Outfit-Statistik hat einen Erwartungswert von 1. Sofern sich die Outfit-Statistik zwischen 0.70 und 1.30 bewegt, 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. Sofern die Infit-Statistik zwischen 0.70 und 1.30 liegt, 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.

10. Literaturverzeichnis

Birnbaum, A. (1958). *On the estimation of mental ability* (Series Report No. 15). Randolph Air Force Base, TX, USA: 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 (Eds.), *Statistical theories of mental test scores* (pp. 425–435). Reading, MA, USA: Addison-Wesley.

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

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

Institut für Bildungsevaluation (2014). *Checks und Aufgabensammlung im Bildungsraum Nordwestschweiz: Pädagogisches Konzept*. Zürich: Institut für Bildungsevaluation. Abgerufen unter https://www.check-dein-wissen.ch/de/konzept/

Kiefer, T., Robitzsch, A., & Wu, M. (2017). TAM: Test analysis modules (Version 1.99999-31) [Software]. Abgerufen unter https://cran.r-project.org/package=TAM.

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

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *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.

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

R Core Team (2016). R: A language and environment for statistical computing [Software]. Wien, Österreich: R Foundation for Statistical Computing. Abgerufen unter https://cran.r-project.org

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.

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

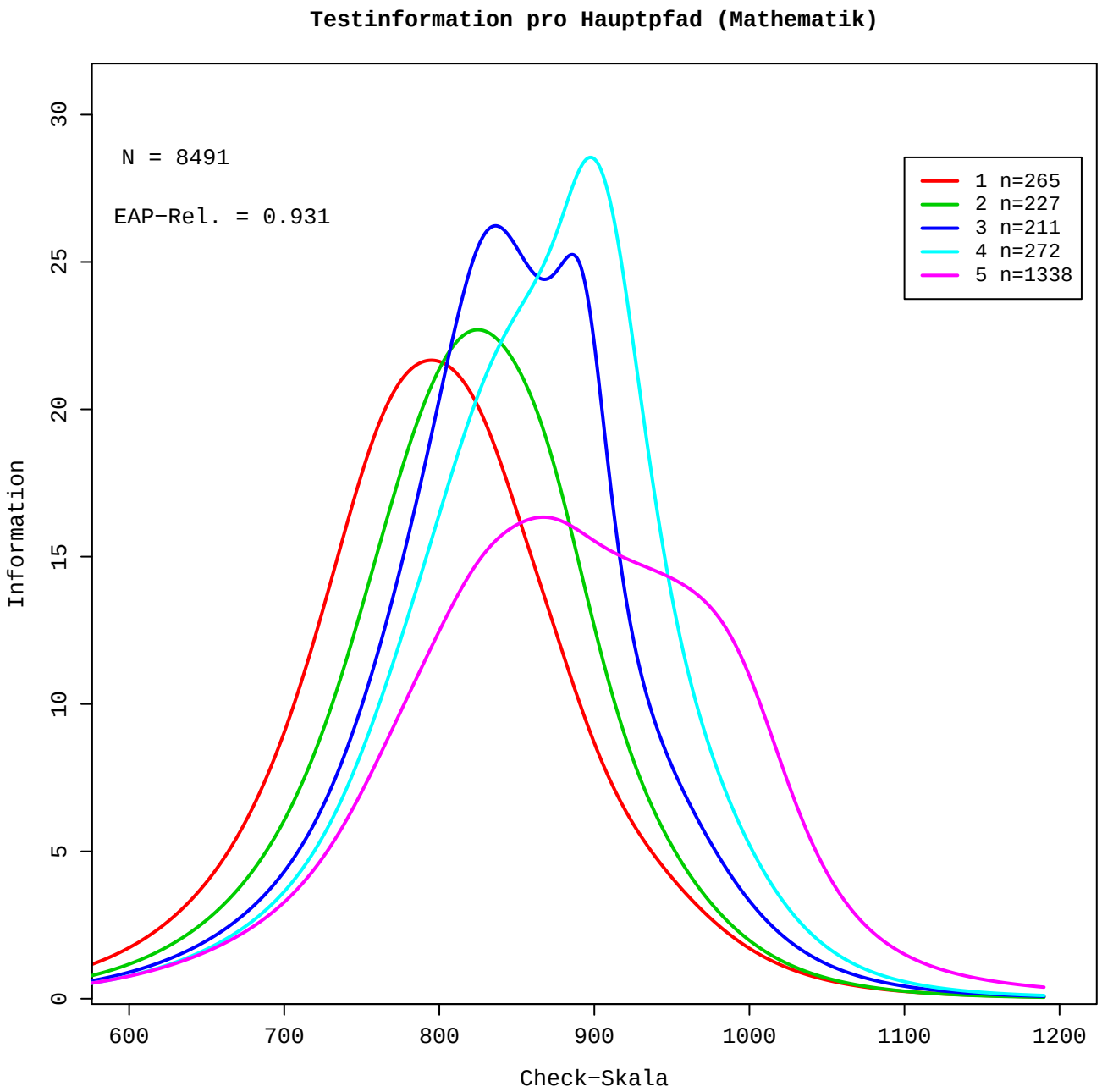
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370–371.

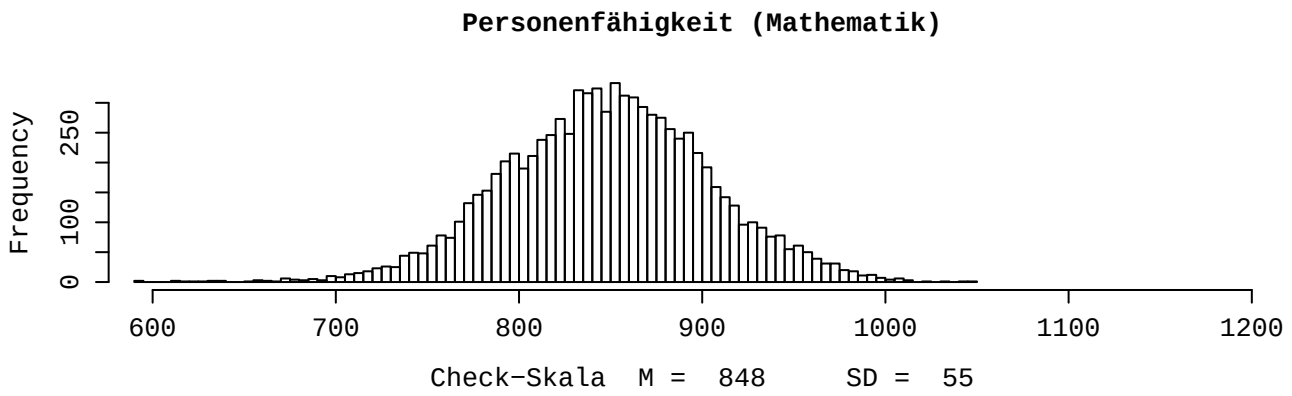
Wright, B. D., & Stone, m. H. (1979). *Best test design. Rasch maesurement.* Chicago: Mesa Press.

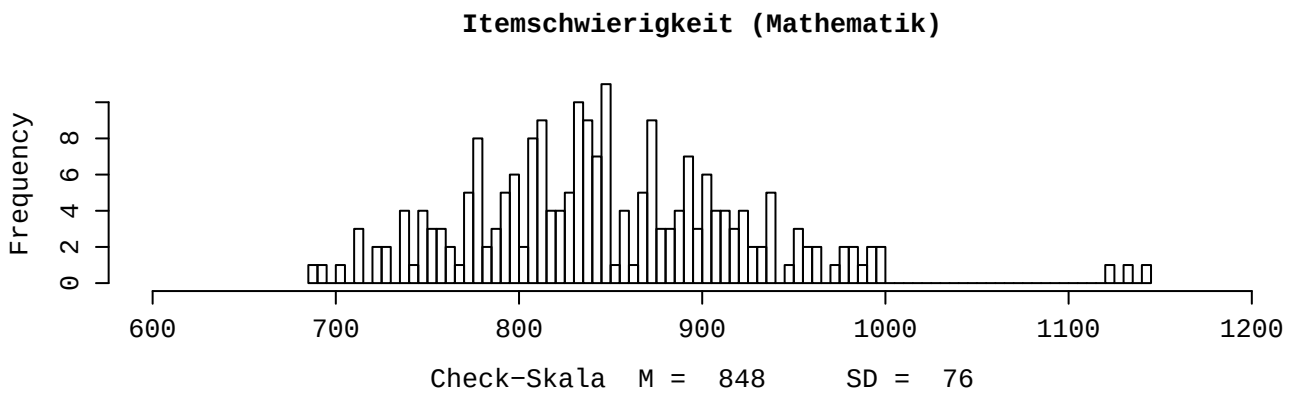
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software manual.* Melbourne, Australia: Australian Council for Educational Research.

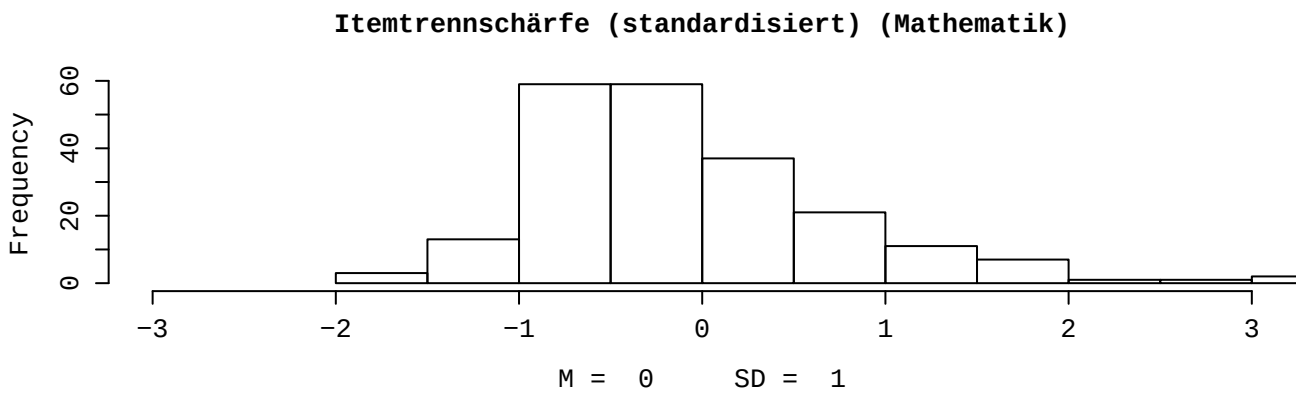
11. Anhang

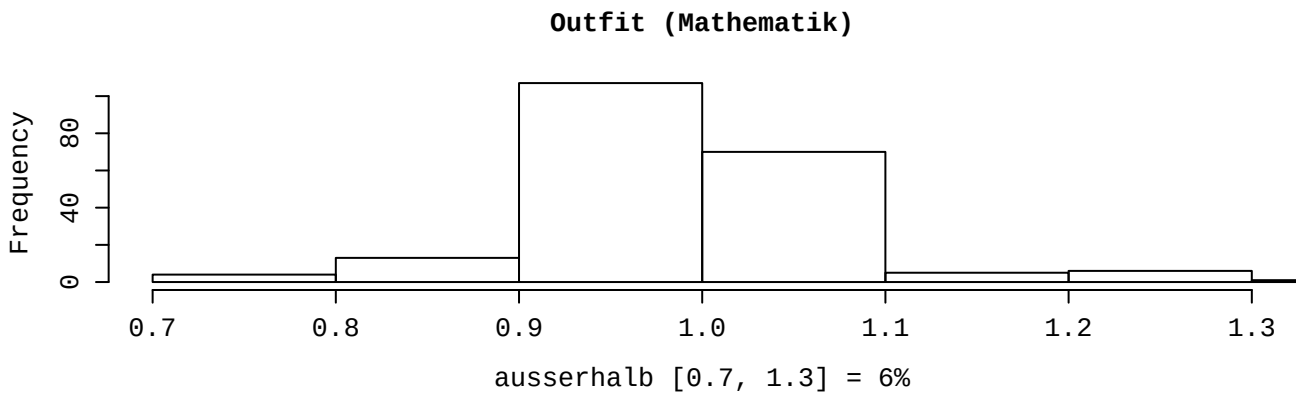
Im Anhang werden pro Skala (Online-Tests) die Testinformation und der Standardfehler nach Hauptpfad (1–5) grafisch dargestellt. Hauptpfade beschreiben hier ein nach Abbildung 1 vertikales Durchlaufen des Tests, wobei Schülerinnen und Schüler Testteile 2 bis 4 auf derselben Schwierigkeitsstufe lösen. Hauptpfad 1 entspricht z.B. einem Pfad, bei dem Schülerinnen und Schüler Testteile 2 bis 4 auf der leichtesten Schwierigkeitsstufe lösen. Der Hauptpfad 5 entspricht demgegenüber einem Pfad, bei dem 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. Für den Online-Test Natur und Technik werden die Testinformation und der Standardfehler für drei statt für fünf Hauptpfade dargestellt. Jeder Pfad beinhaltet die Information des obligatorischen Testteils „Allgemeine Naturwissenschaften“ sowie von zwei frei gewählten Testteilen aus den Kompetenzbereichen Biologie, Chemie und Physik. Hauptpfad 1 bedeutet, dass bei beiden frei wählbaren Subtests die einfache Schwierigkeitsstufe gelöst wird, Hauptpfad 3 bedeutet, dass bei beiden Subtests die schwierigere Stufe gelöst wird. Hauptpfad 2 bedeutet schliesslich, dass bei einem der beiden Subtests die einfache und beim anderen die schwierigere Stufe gelöst wird. Im Anhang werden zum Schluss ebenfalls Tabellen mit Skalenmittelwerten und Standardabweichungen auf der Check-Skala sowie Skalen-Interkorrelationen ausgewiesen.

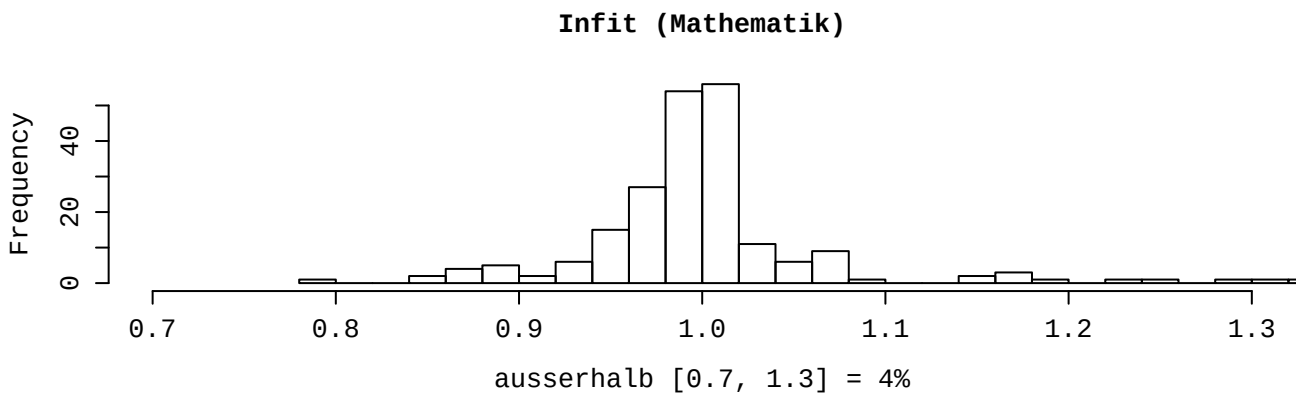

Testinformation pro Hauptpfad (Mathematik)
N = 8491
EAP-Rel. = 0.931
1 n=265
2 n=227
3 n=211
4 n=272
5 n=1338
Information
0
5
10
15
20
25
30
600
700
800
900
1000
1100
1200
Check-Skala

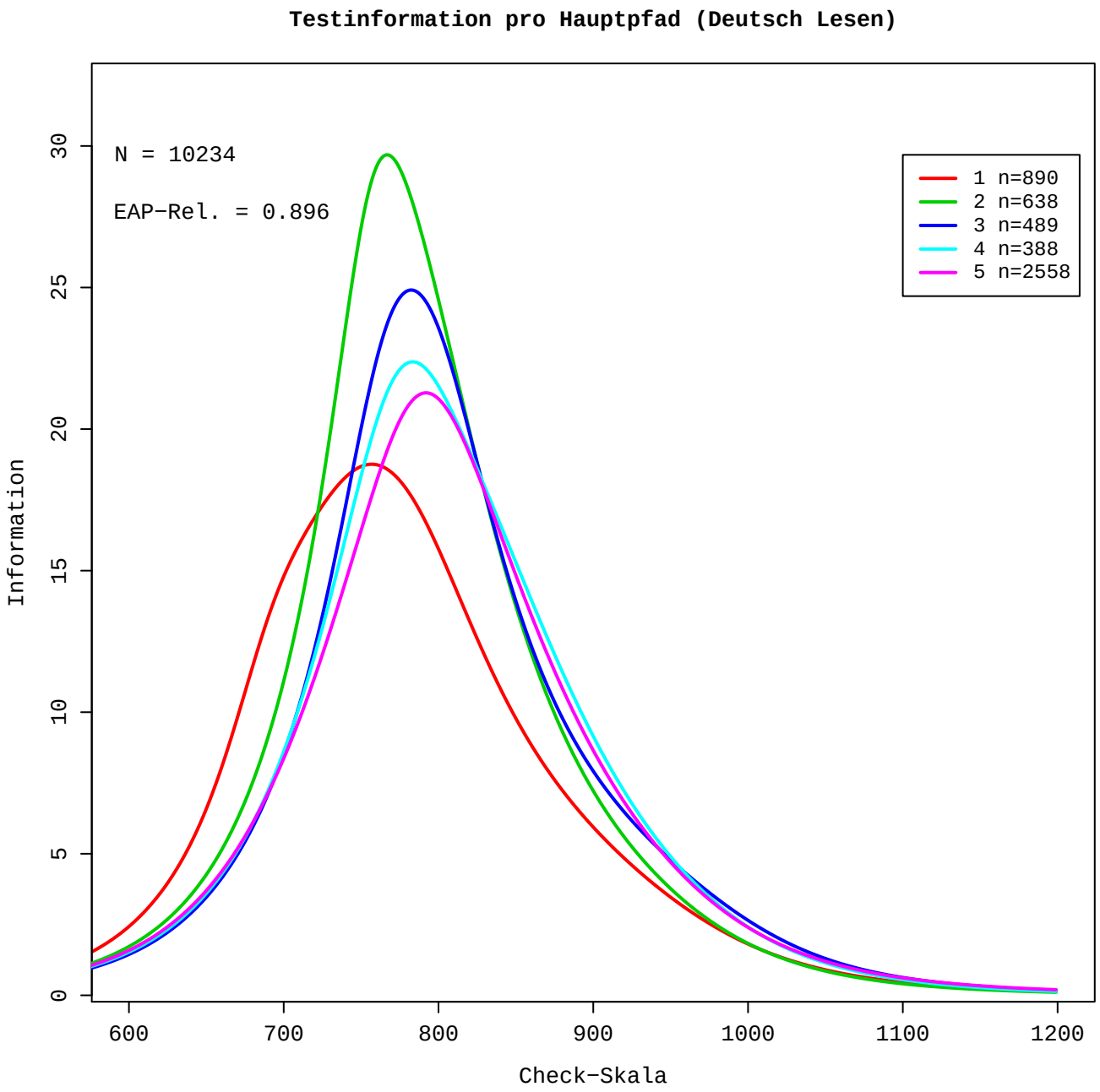

Personenfähigkeit (Mathematik)
Frequency
0
100
250
600
700
800
900
1000
1100
1200
Check−Skala M = 848 SD = 55

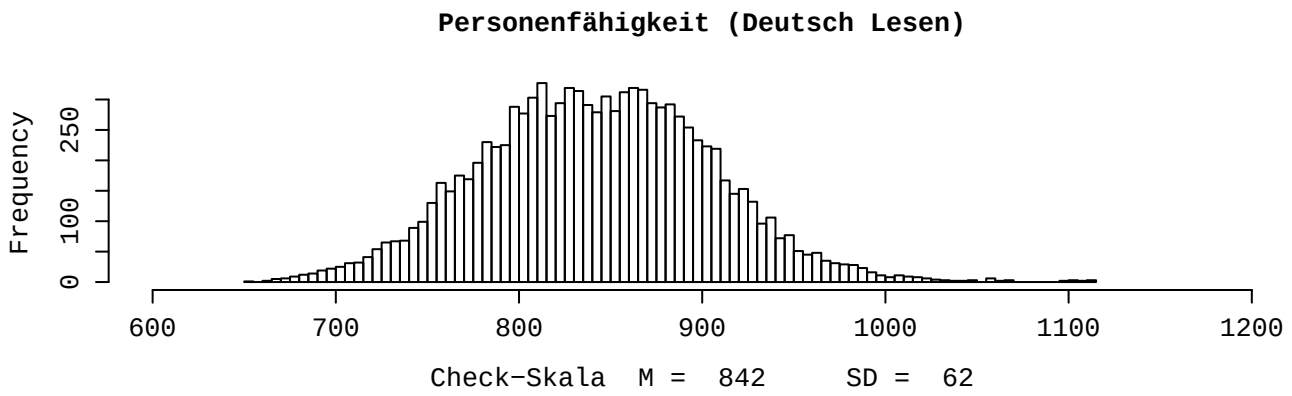

Itemschwierigkeit (Mathematik)
Frequency
0
2
4
6
8
600
700
800
900
1000
1100
1200
Check−Skala M = 848 SD = 76

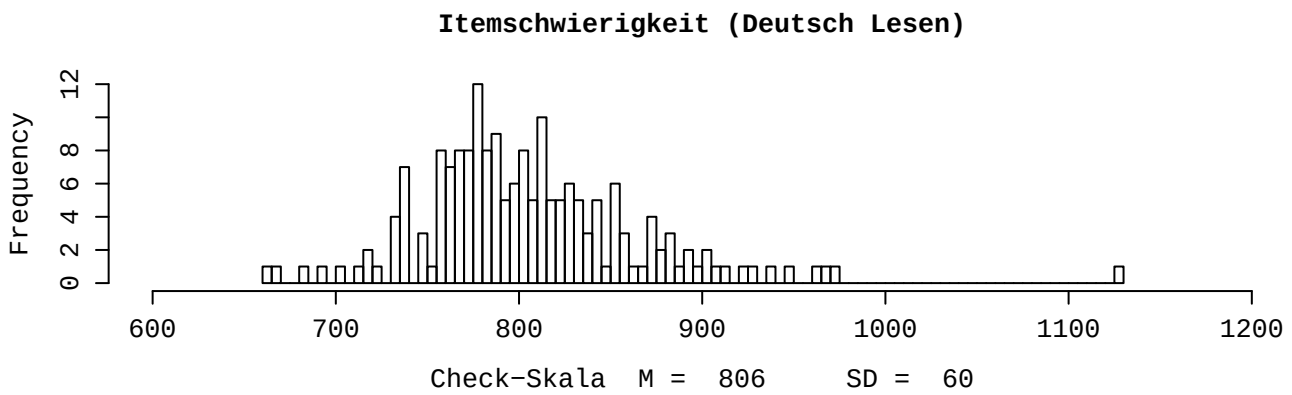

Itemtrennschärfe (standardisiert) (Mathematik)
Frequency
0
20
40
60
−3
−2
−1
0
1
2
3
M = 0 SD = 1

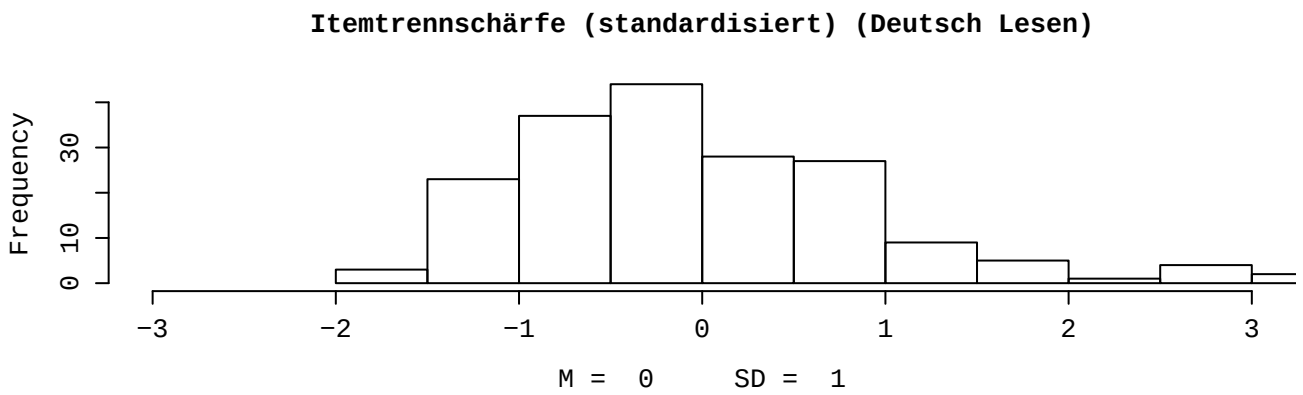

Outfit (Mathematik)
Frequency
0
40
80
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 6%

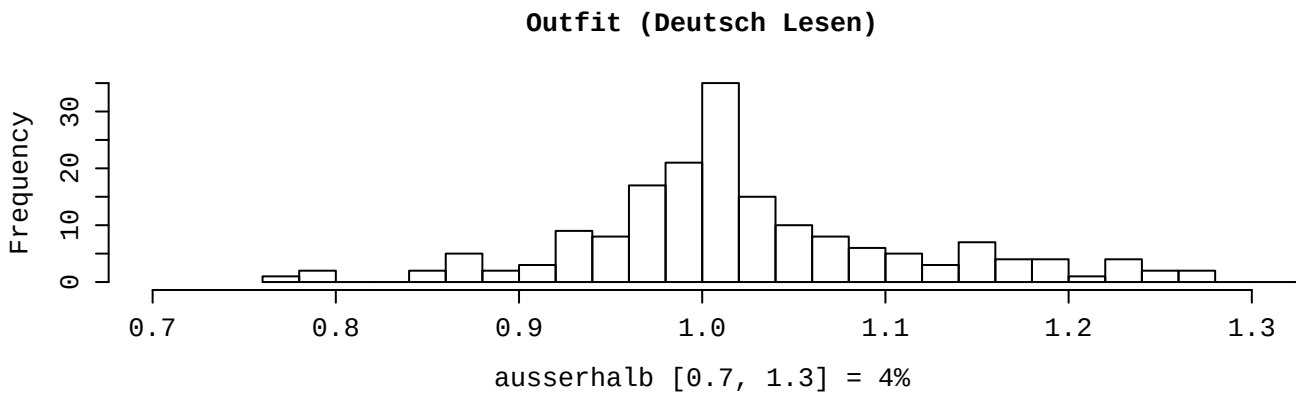

Infit (Mathematik)
Frequency
0
20
40
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 4%

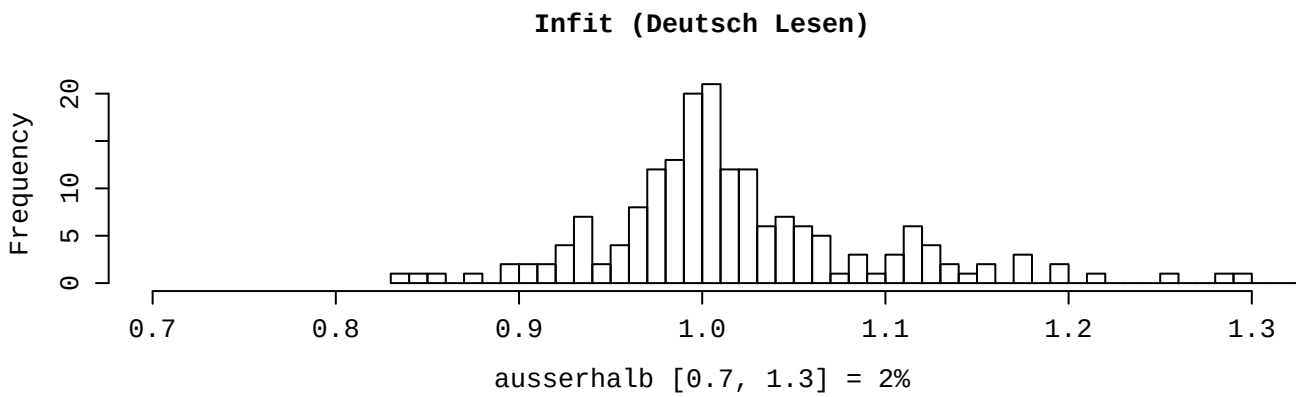

Testinformation pro Hauptpfad (Deutsch Lesen)
N = 10234
EAP-Rel. = 0.896
1 n=890
2 n=638
3 n=489
4 n=388
5 n=2558
Information
0
5
10
15
20
25
30
600
700
800
900
1000
1100
1200
Check-Skala

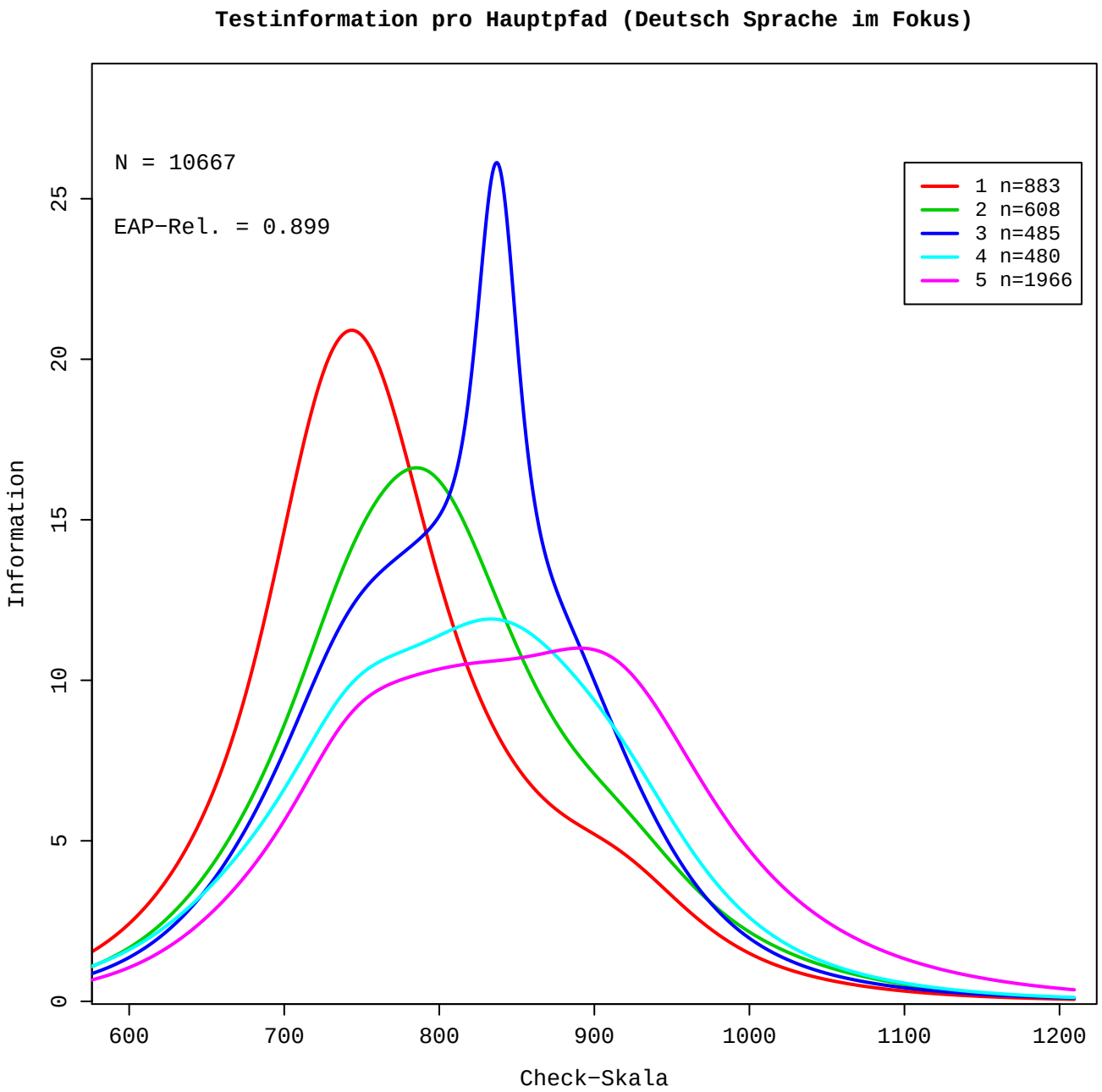

Personenfähigkeit (Deutsch Lesen)
Frequency
0
100
250
600
700
800
900
1000
1100
1200
Check−Skala M = 842 SD = 62

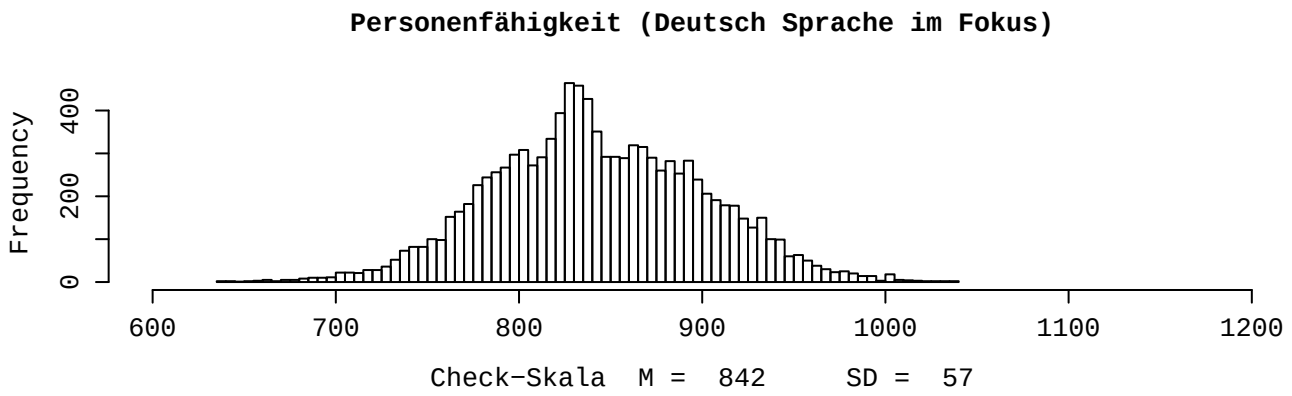

Itemschwierigkeit (Deutsch Lesen)
Frequency
0
2
4
6
8
12
600
700
800
900
1000
1100
1200
Check−Skala M = 806 SD = 60

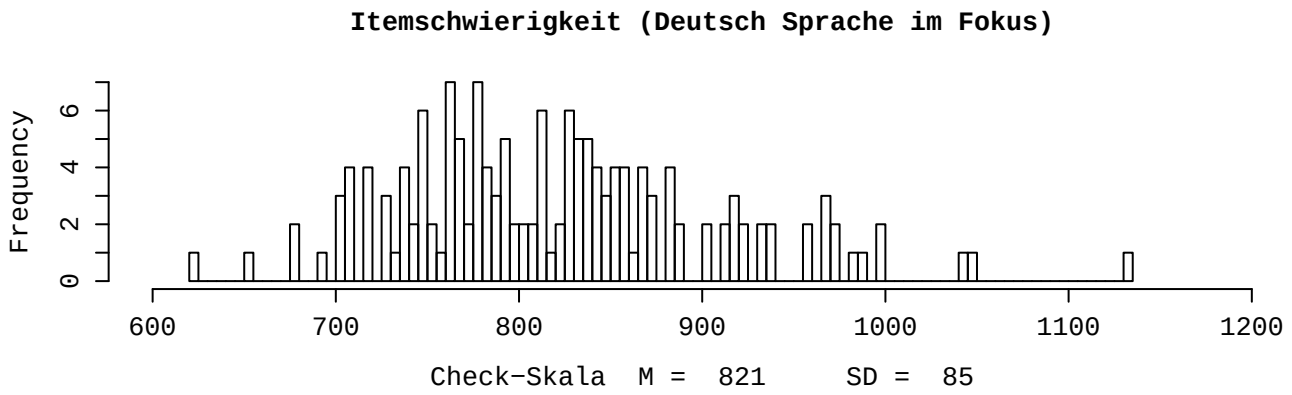

Itemtrennschärfe (standardisiert) (Deutsch Lesen)
Frequency
0
10
30
−3
−2
−1
0
1
2
3
M = 0 SD = 1

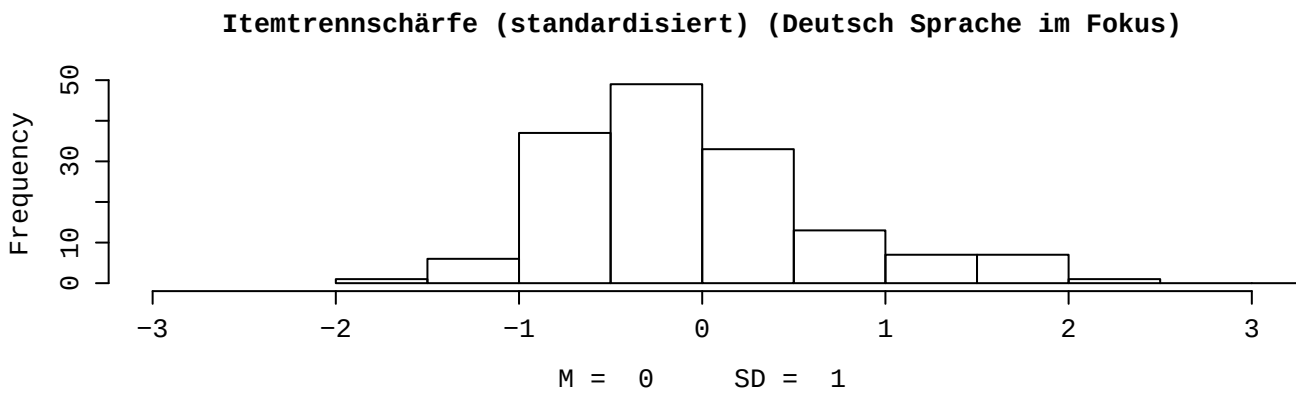

Outfit (Deutsch Lesen)
Frequency
0
10
20
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 4%

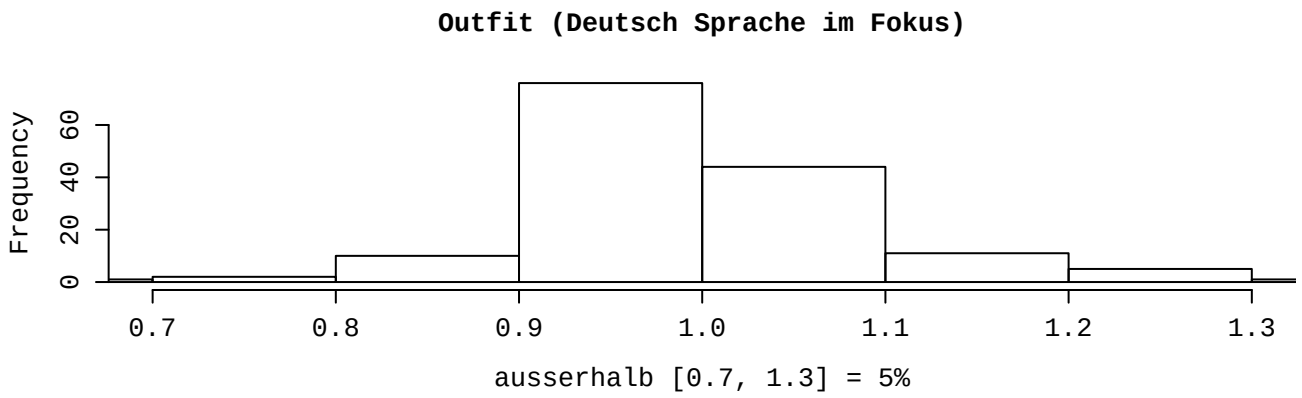

Infit (Deutsch Lesen)
Frequency
0
5
10
20
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 2%

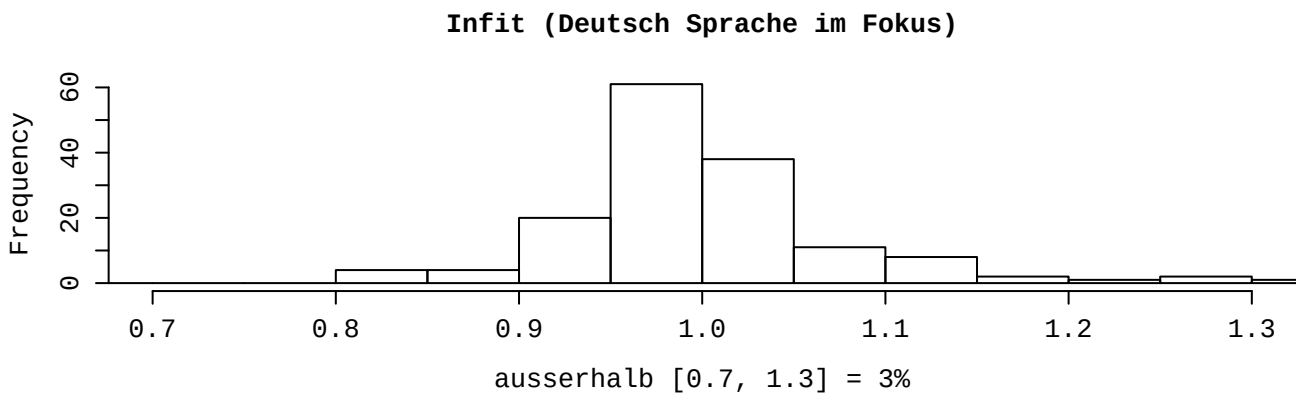

Testinformation pro Hauptpfad (Deutsch Sprache im Fokus)
N = 10667
EAP-Rel. = 0.899
1 n=883
2 n=608
3 n=485
4 n=480
5 n=1966
Information
0
5
10
15
20
25
600
700
800
900
1000
1100
1200
Check-Skala

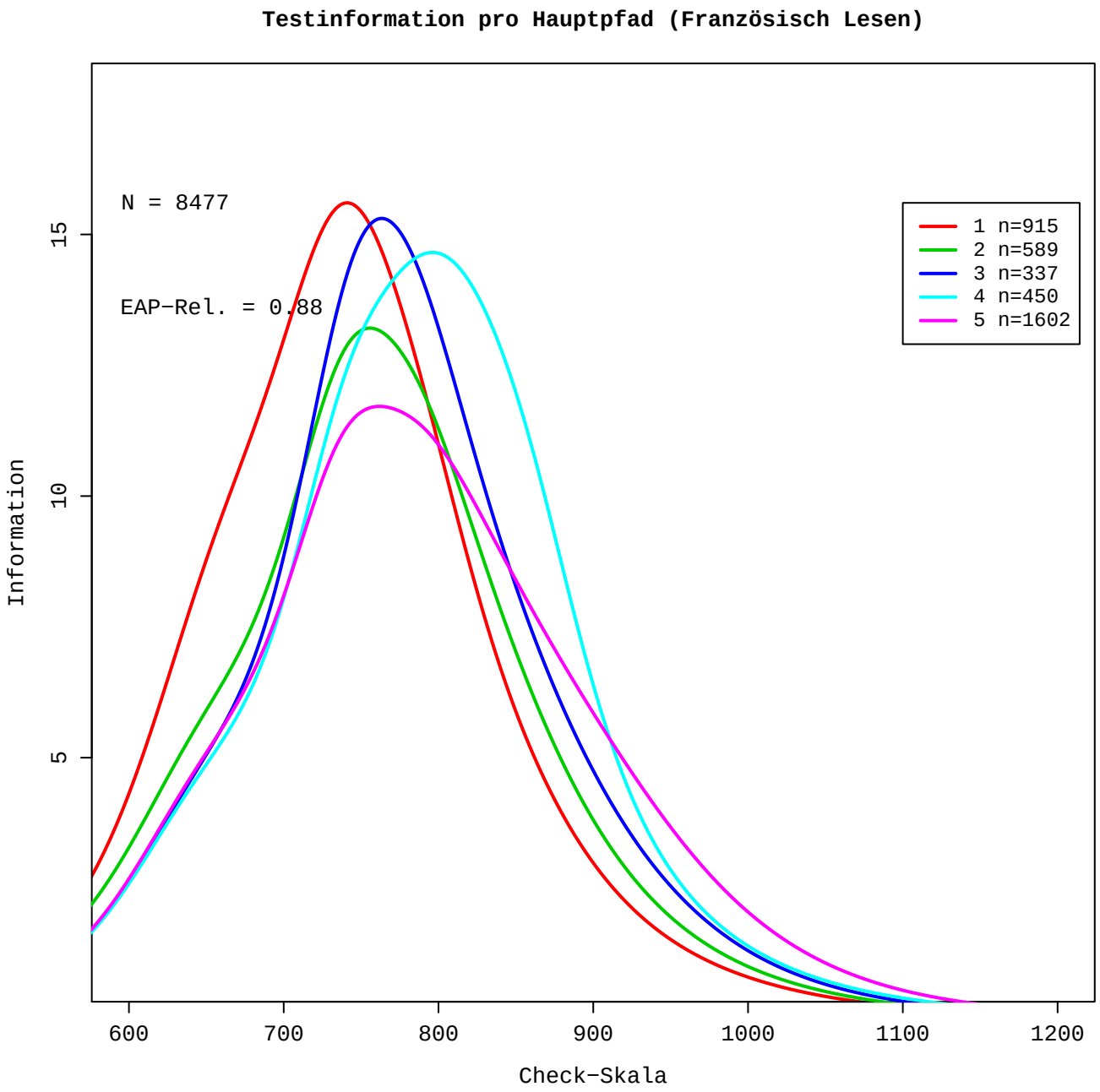

Personenfähigkeit (Deutsch Sprache im Fokus)
Frequency
0
200
400
600
700
800
900
1000
1100
1200
Check−Skala M = 842 SD = 57

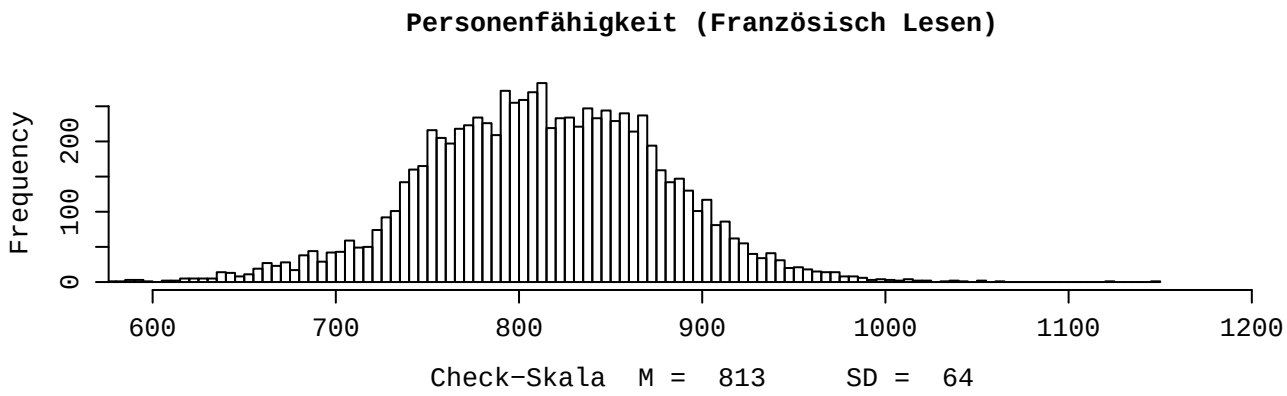

Itemschwierigkeit (Deutsch Sprache im Fokus)
Frequency
0
2
4
6
600
700
800
900
1000
1100
1200
Check−Skala M = 821 SD = 85

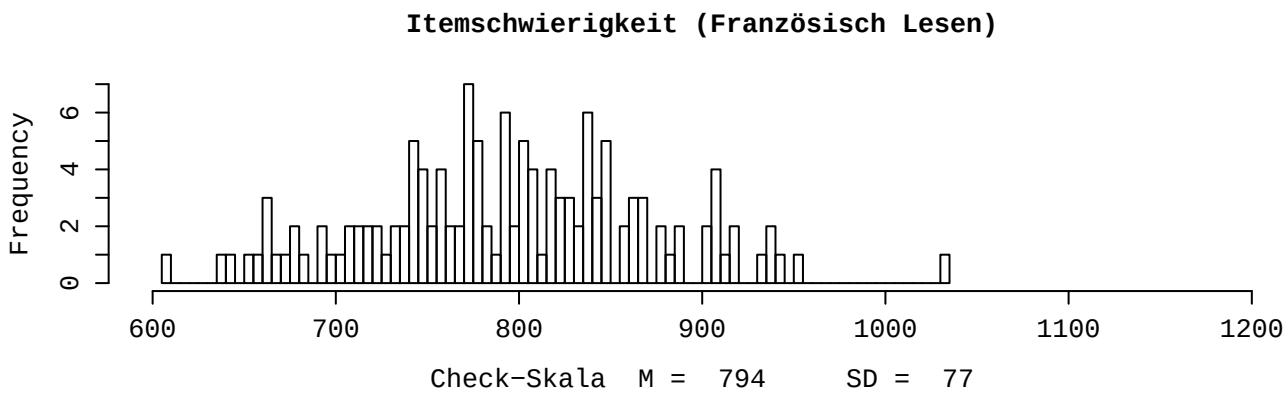

Itemtrennschärfe (standardisiert) (Deutsch Sprache im Fokus)
Frequency
0
10
30
50
−3
−2
−1
0
1
2
3
M = 0 SD = 1

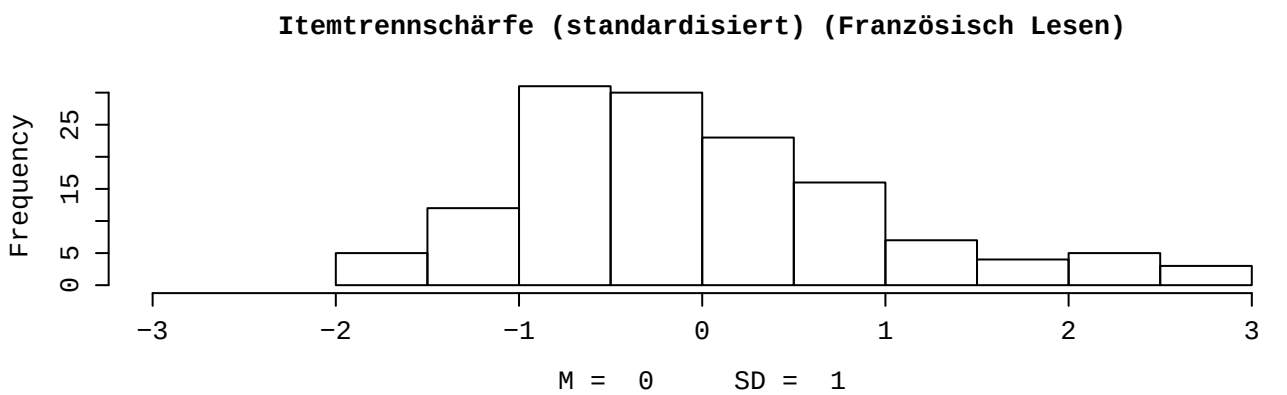

Outfit (Deutsch Sprache im Fokus)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 5%

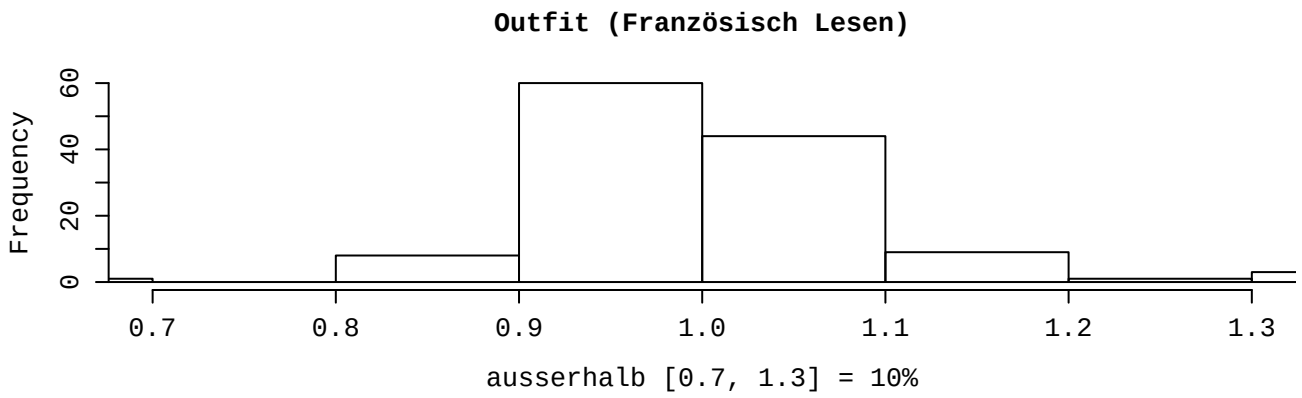

Infit (Deutsch Sprache im Fokus)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 3%

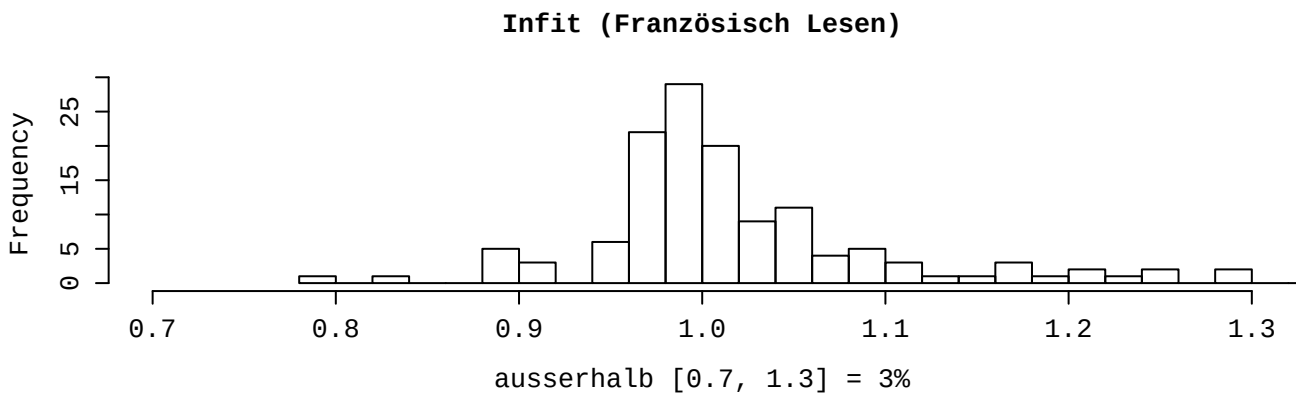

Testinformation pro Hauptpfad (Französisch Lesen)
N = 8477
EAP−Rel. = 0.88
1 n=915
2 n=589
3 n=337
4 n=450
5 n=1602
Information
5
10
15
600
700
800
900
1000
1100
1200
Check−Skala

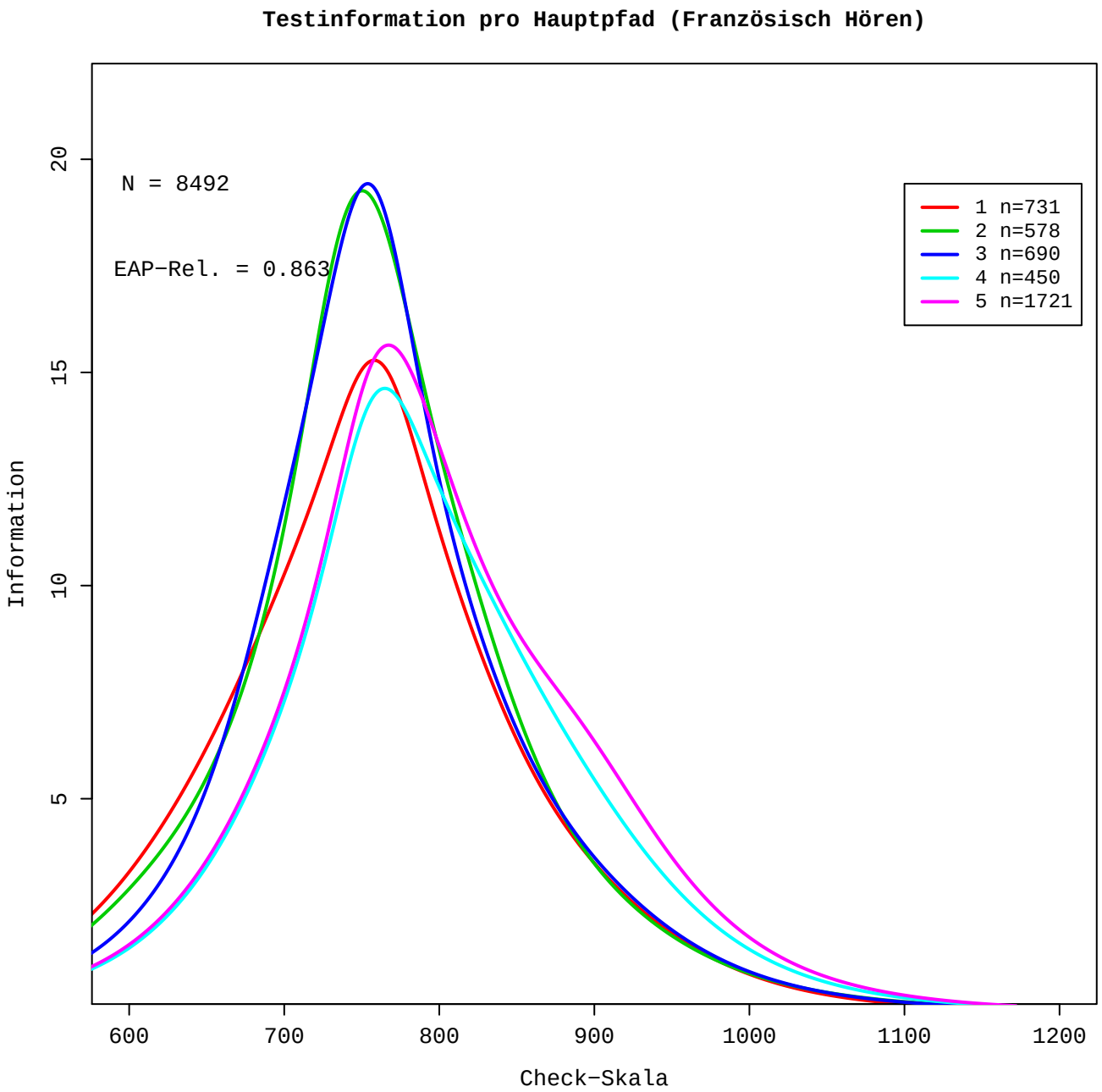

Personenfähigkeit (Französisch Lesen)
Frequency
0
100
200
600
700
800
900
1000
1100
1200
Check−Skala M = 813 SD = 64

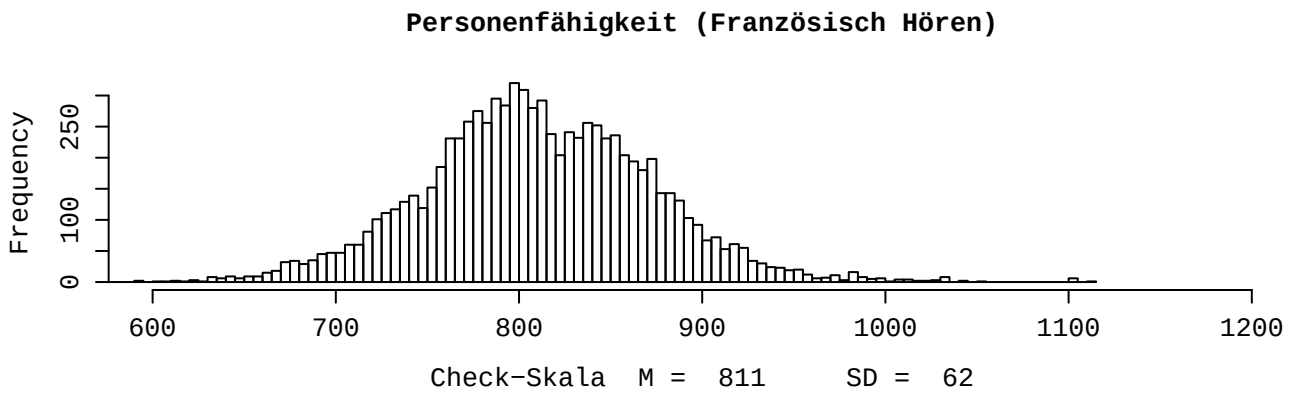

Itemschwierigkeit (Französisch Lesen)
Frequency
0
2
4
6
600
700
800
900
1000
1100
1200
Check−Skala M = 794 SD = 77

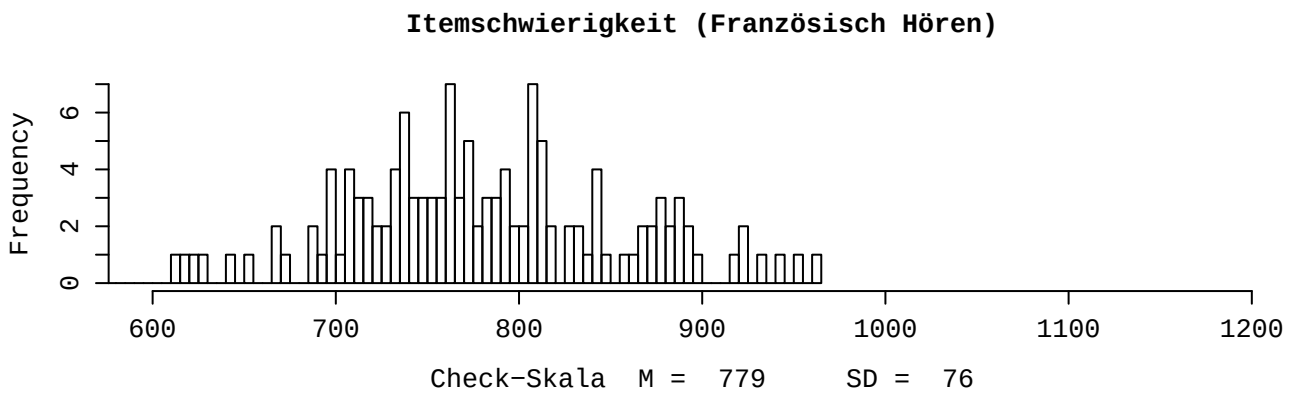

Itemtrennschärfe (standardisiert) (Französisch Lesen)
Frequency
0
5
15
25
−3
−2
−1
0
1
2
3
M = 0 SD = 1

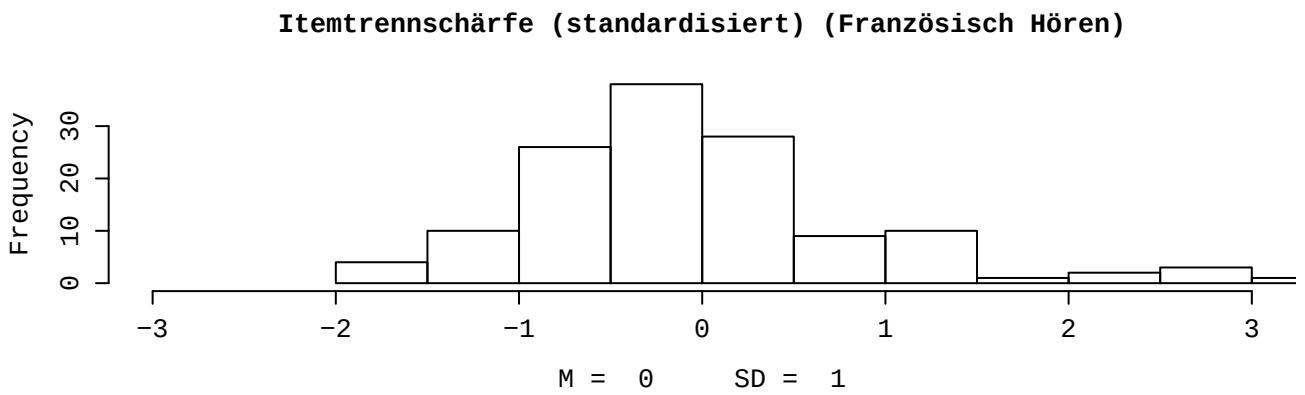

Outfit (Französisch Lesen)
Frequency
0
20
40
60
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 10%

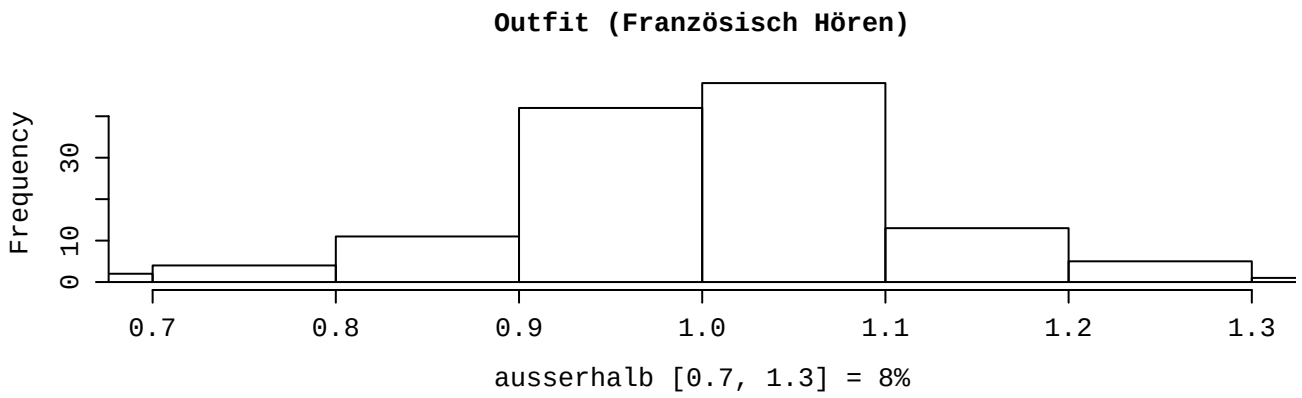

Infit (Französisch Lesen)
Frequency
0
5
15
25
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 3%

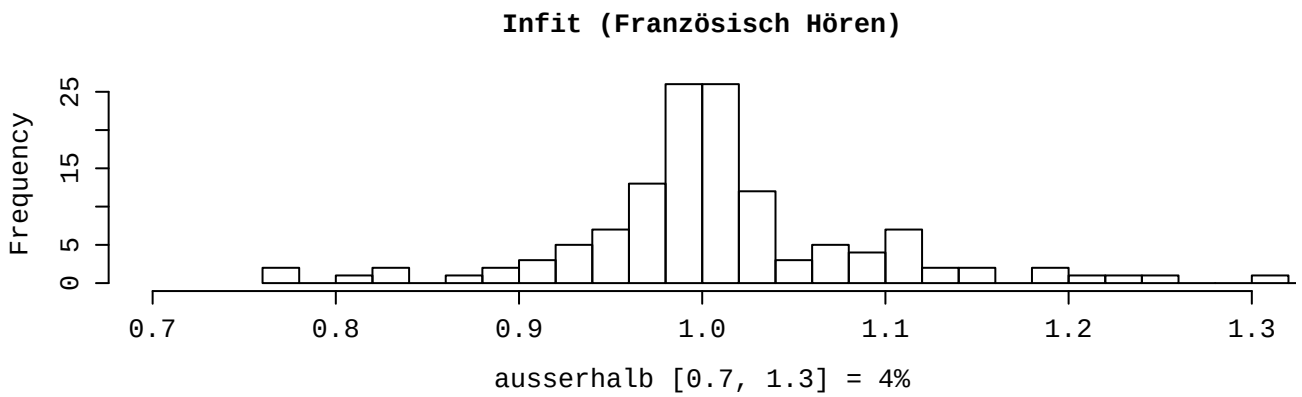

Testinformation pro Hauptpfad (Französisch Hören)
N = 8492
EAP-Rel. = 0.863
1 n=731
2 n=578
3 n=690
4 n=450
5 n=1721
Information
5
10
15
20
600
700
800
900
1000
1100
1200
Check-Skala

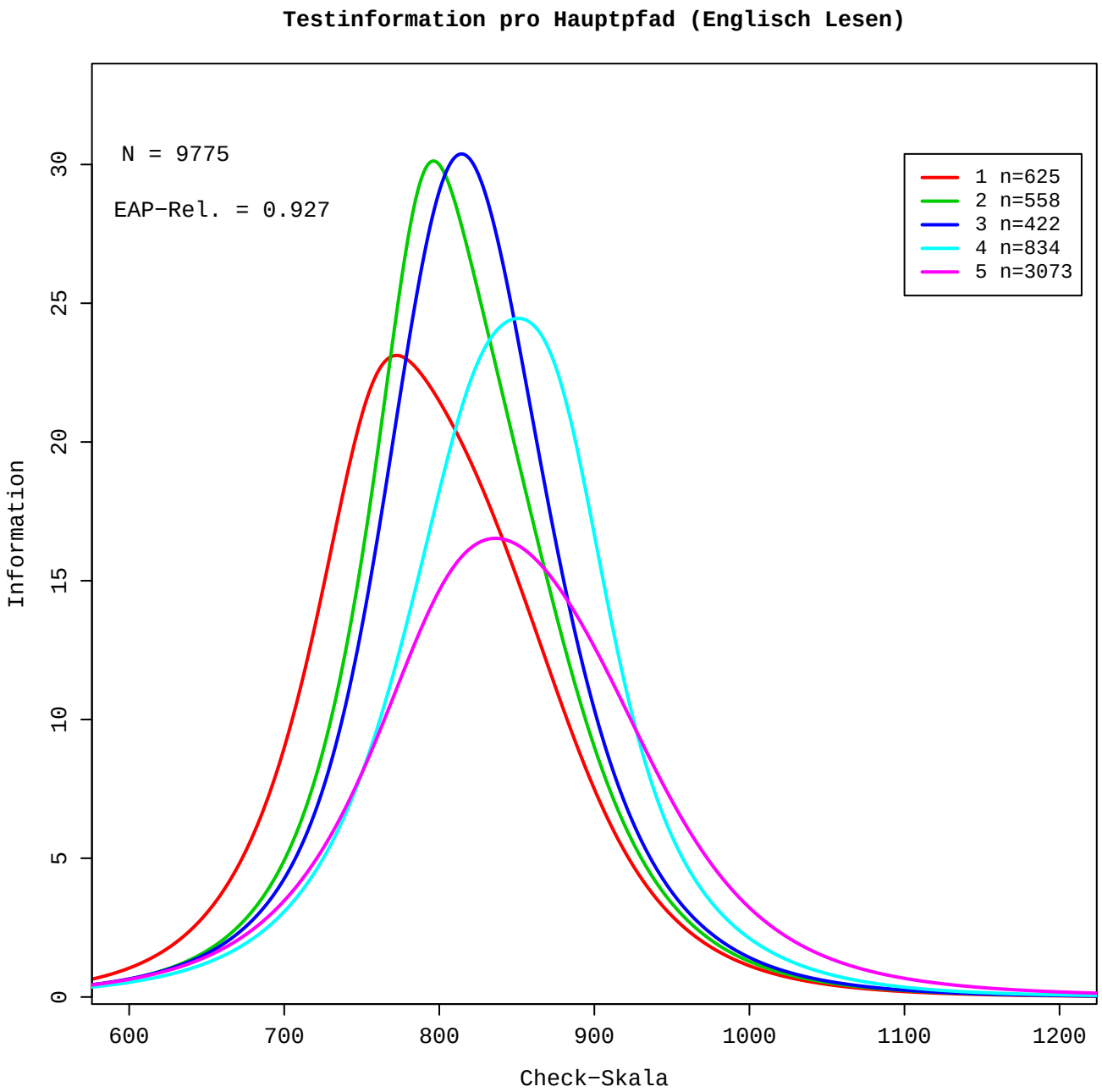

Personenfähigkeit (Französisch Hören)
Frequency
0
100
250
600
700
800
900
1000
1100
1200
Check−Skala M = 811 SD = 62

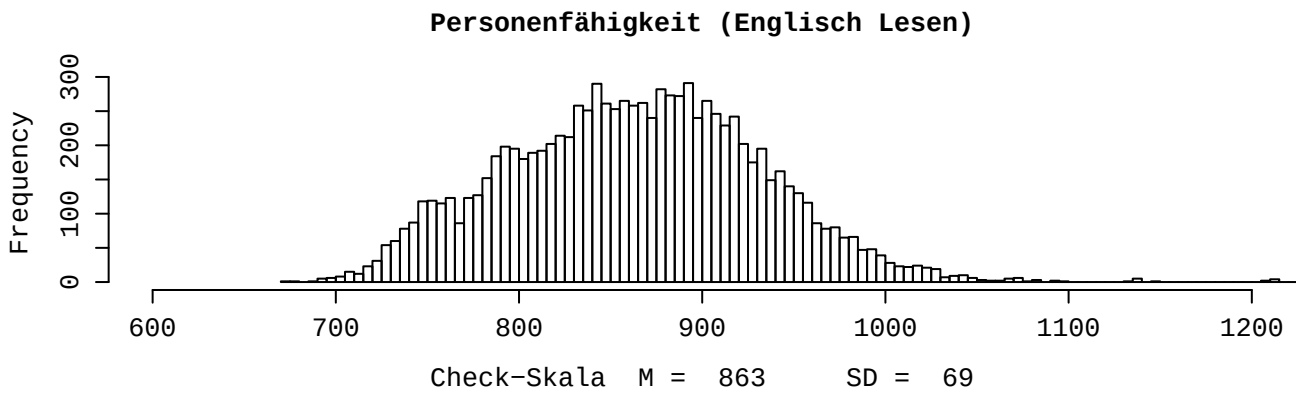

Itemschwierigkeit (Französisch Hören)
Frequency
0
2
4
6
600
700
800
900
1000
1100
1200
Check−Skala M = 779 SD = 76

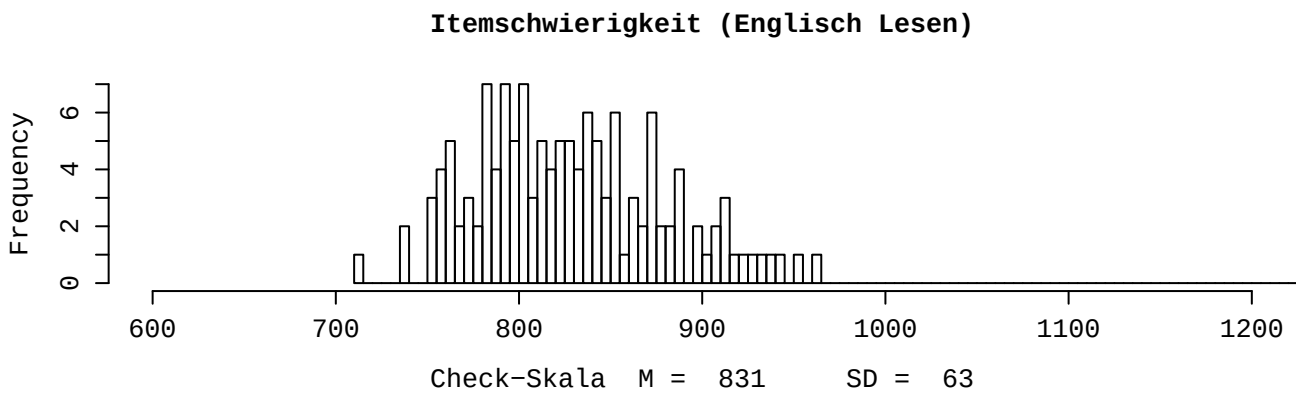

Itemtrennschärfe (standardisiert) (Französisch Hören)
Frequency
0
10
20
30
−3
−2
−1
0
1
2
3
M = 0 SD = 1

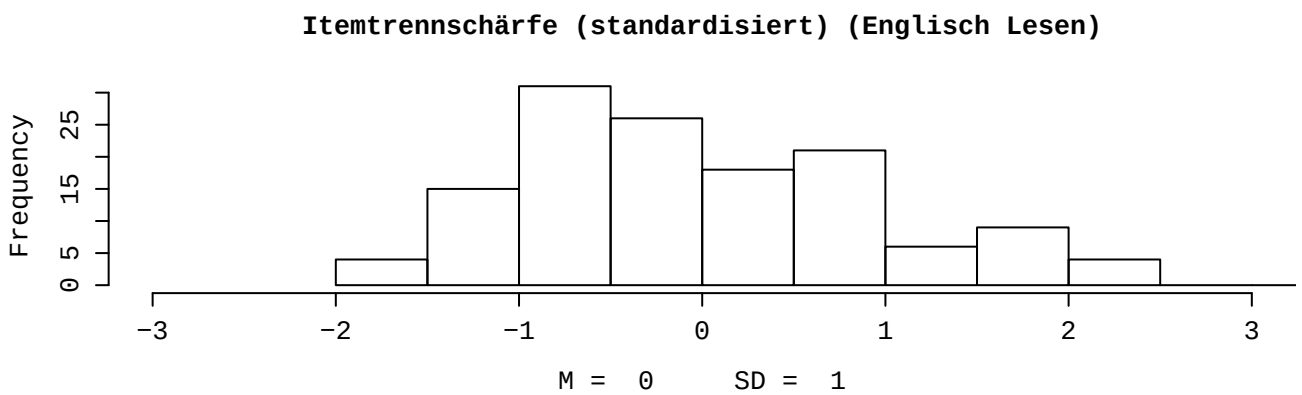

Outfit (Französisch Hören)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 8%

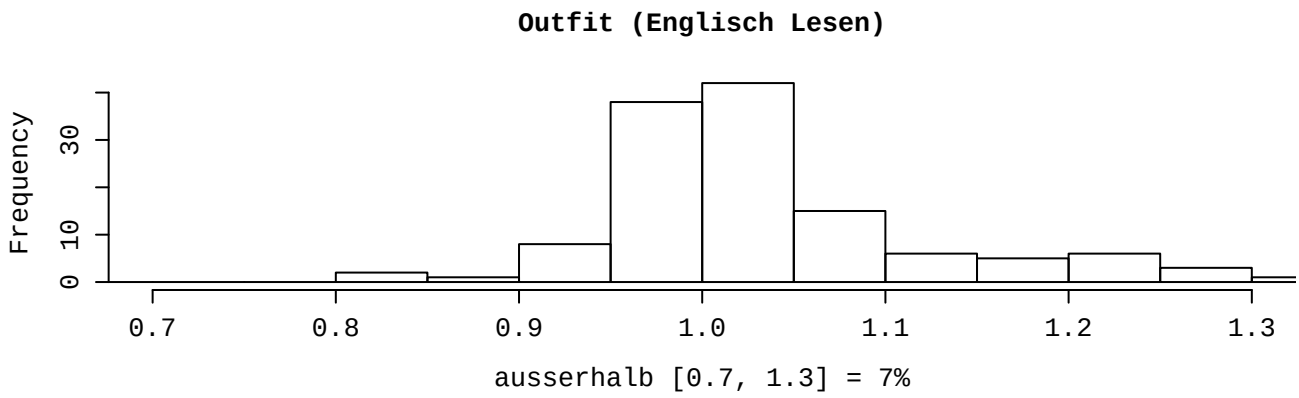

Infit (Französisch Hören)
Frequency
0
5
15
25
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 4%

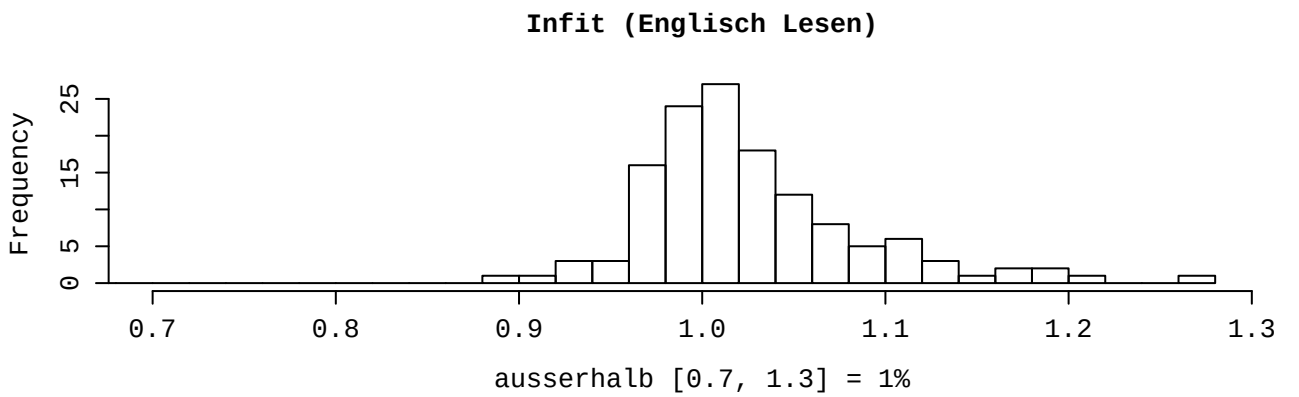

Testinformation pro Hauptpfad (Englisch Lesen)
N = 9775
EAP-Rel. = 0.927
1 n=625
2 n=558
3 n=422
4 n=834
5 n=3073
Information
0
5
10
15
20
25
30
600
700
800
900
1000
1100
1200
Check-Skala

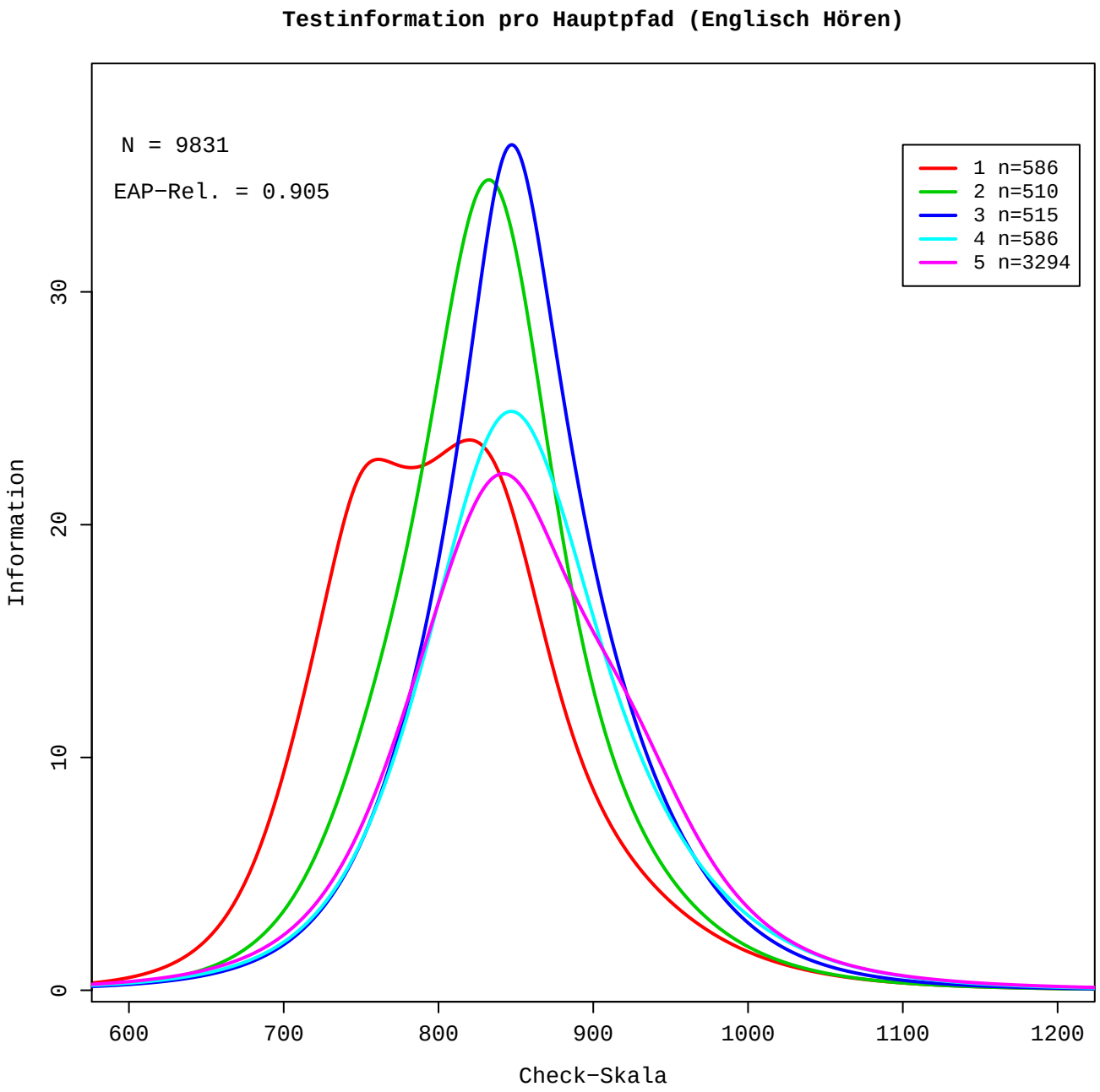

Personenfähigkeit (Englisch Lesen)
Frequency
0
100
200
300
600
700
800
900
1000
1100
1200
Check−Skala M = 863 SD = 69

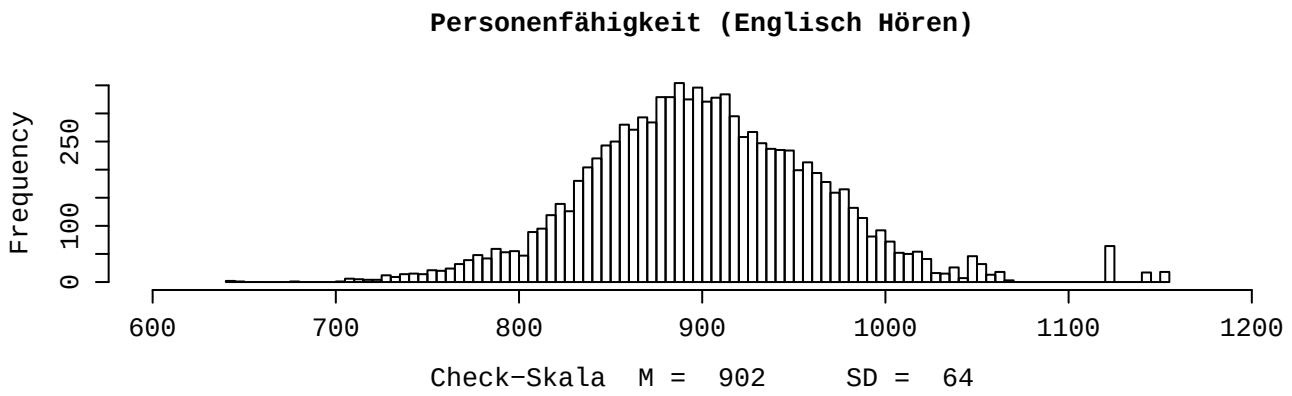

Itemschwierigkeit (Englisch Lesen)
Frequency
0
2
4
6
600
700
800
900
1000
1100
1200
Check−Skala M = 831 SD = 63

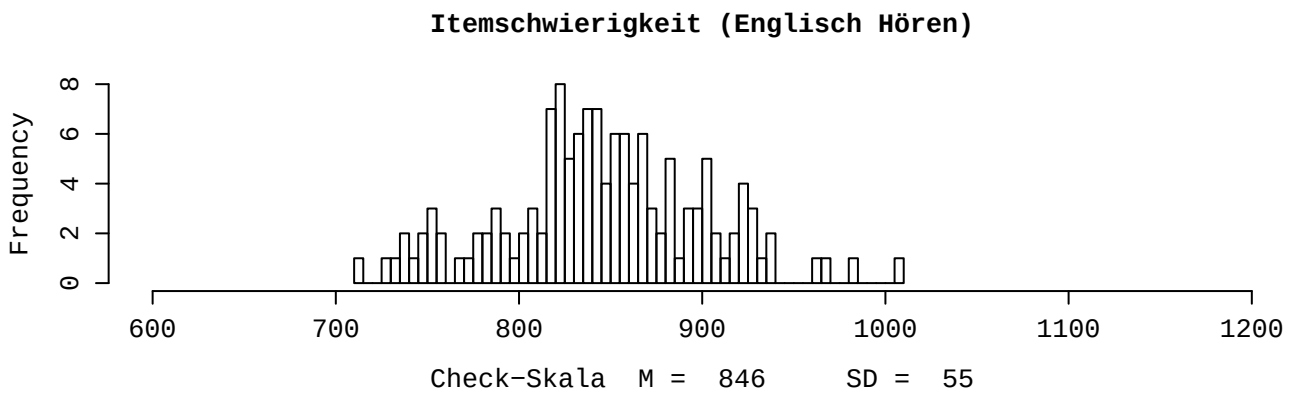

Itemtrennschärfe (standardisiert) (Englisch Lesen)
Frequency
0
5
15
25
−3
−2
−1
0
1
2
3
M = 0 SD = 1

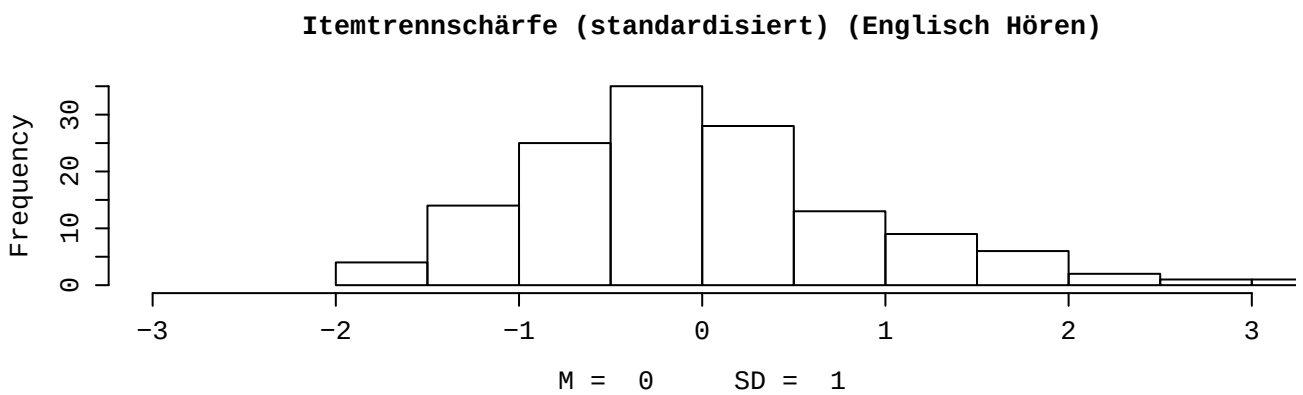

Outfit (Englisch Lesen)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 7%

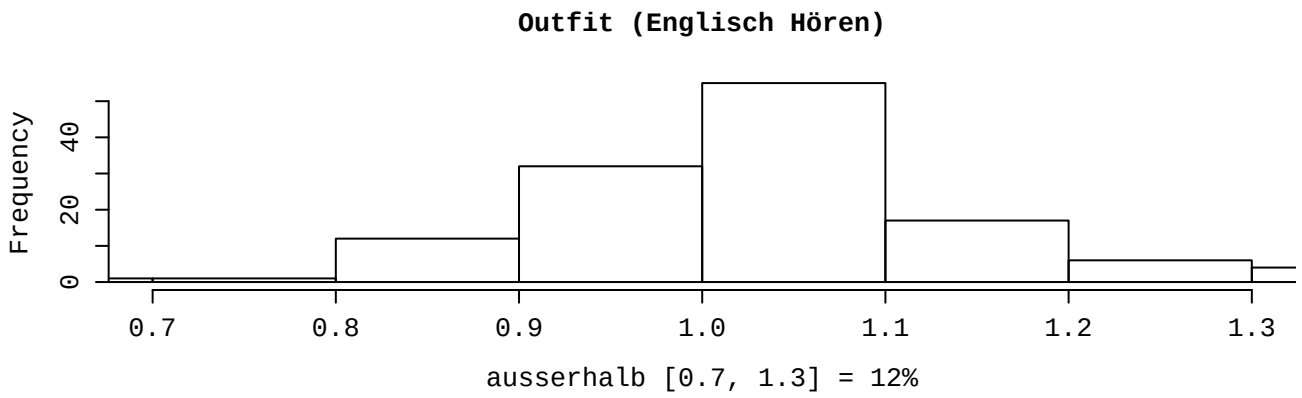

Infit (Englisch Lesen)
Frequency
0
5
15
25
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 1%

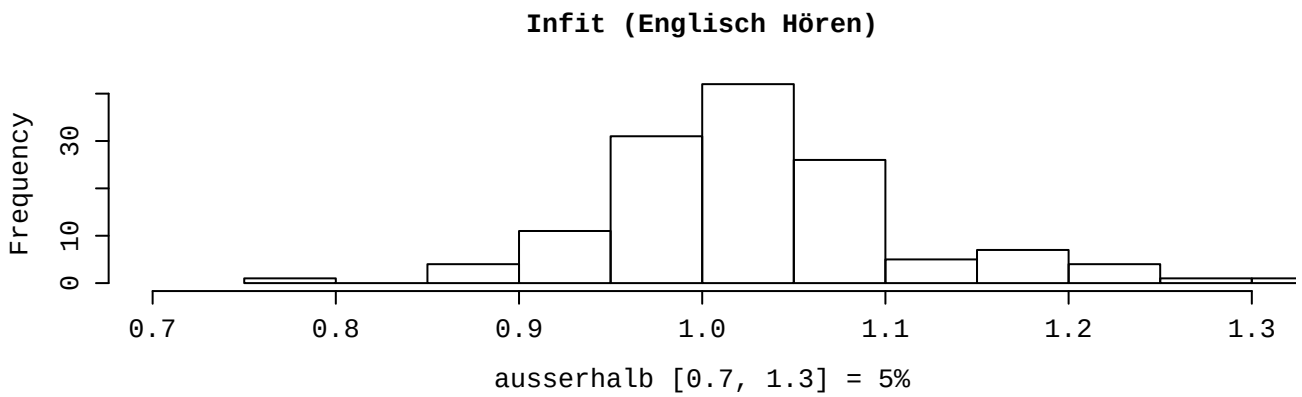

Testinformation pro Hauptpfad (Englisch Hören)
N = 9831
EAP-Rel. = 0.905
1 n=586
2 n=510
3 n=515
4 n=586
5 n=3294
Information
0
10
20
30
600
700
800
900
1000
1100
1200
Check-Skala

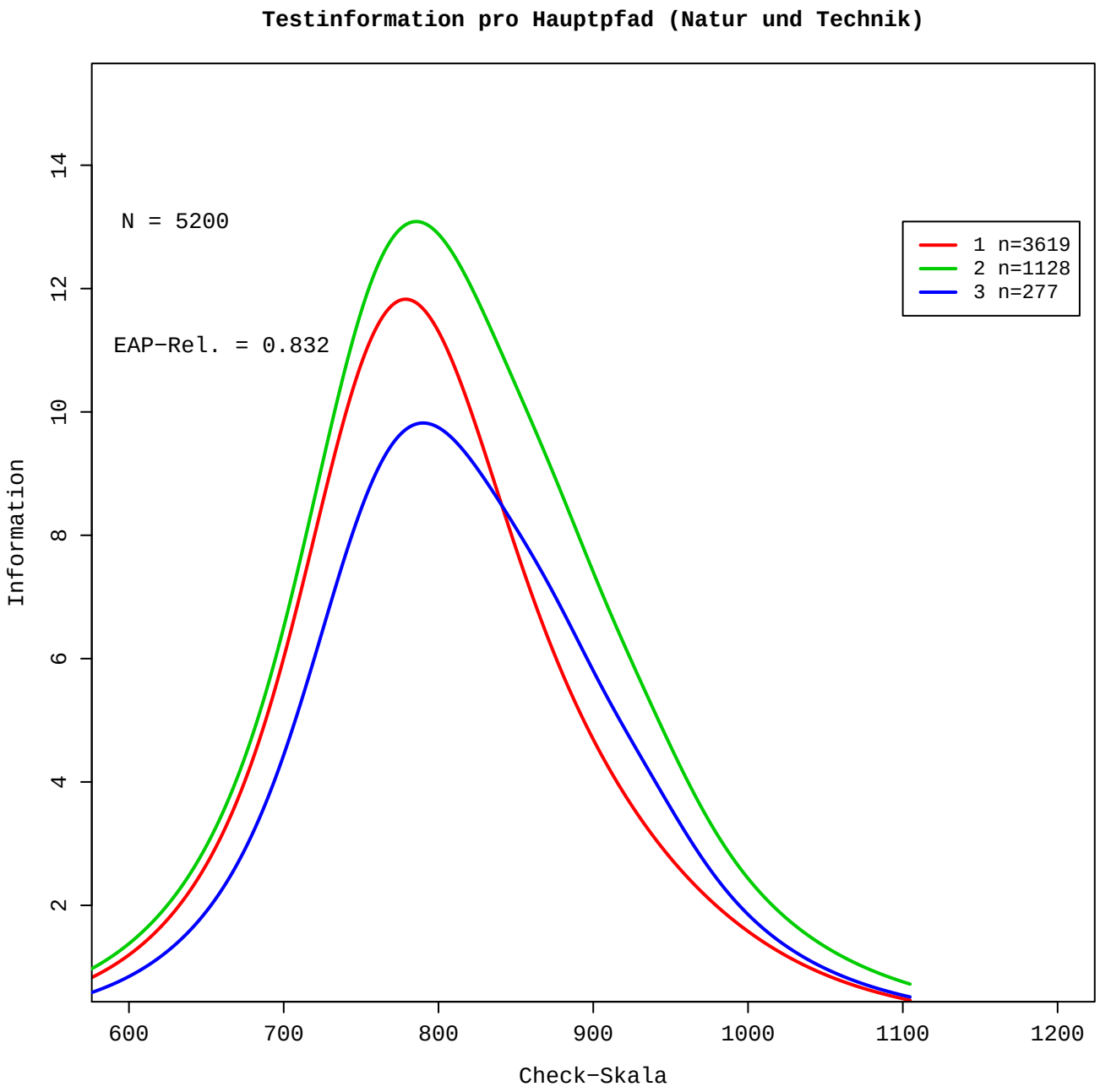

Personenfähigkeit (Englisch Hören)
Frequency
0
100
250
600
700
800
900
1000
1100
1200
Check−Skala M = 902 SD = 64

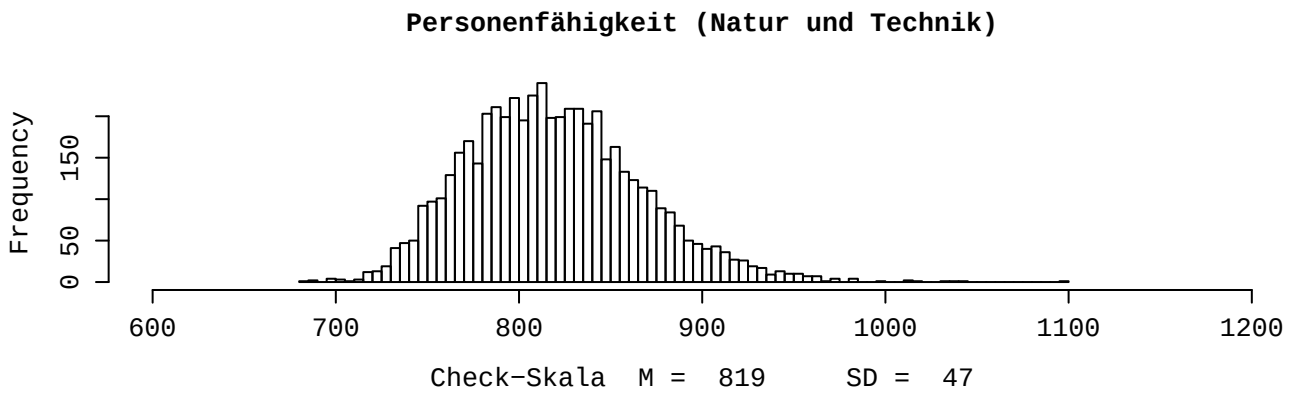

Itemschwierigkeit (Englisch Hören)
Frequency
0
2
4
6
8
600
700
800
900
1000
1100
1200
Check−Skala M = 846 SD = 55

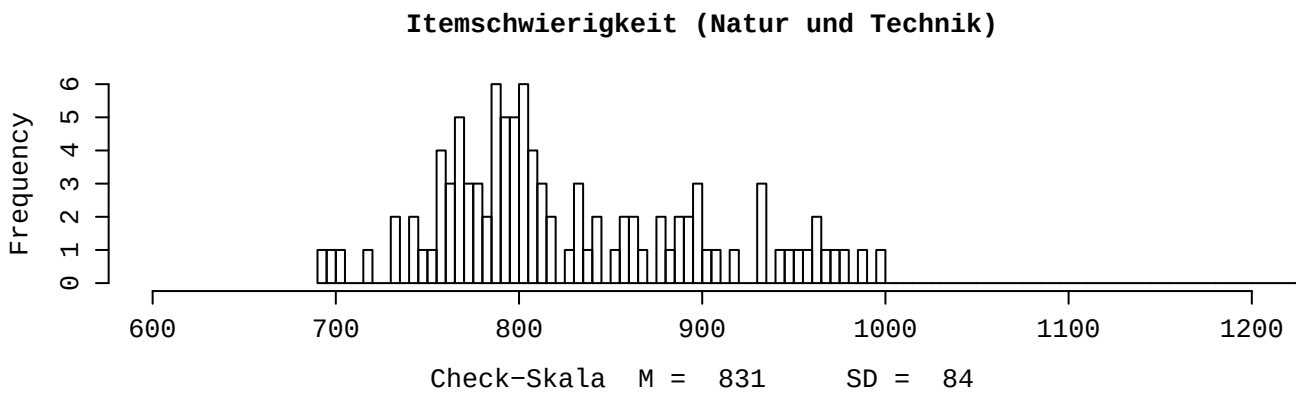

Itemtrennschärfe (standardisiert) (Englisch Hören)
Frequency
0
10
20
30
−3
−2
−1
0
1
2
3
M = 0 SD = 1

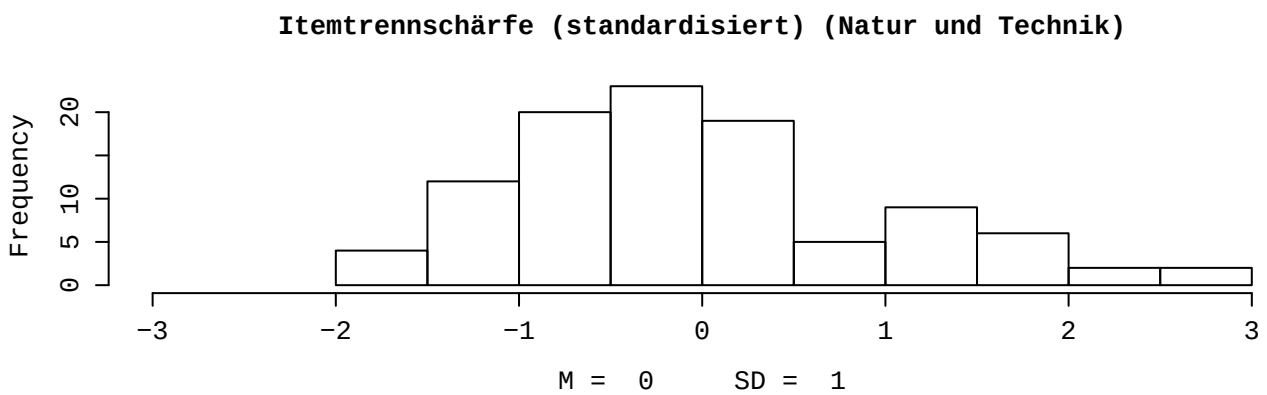

Outfit (Englisch Hören)
Frequency
0
20
40
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 12%

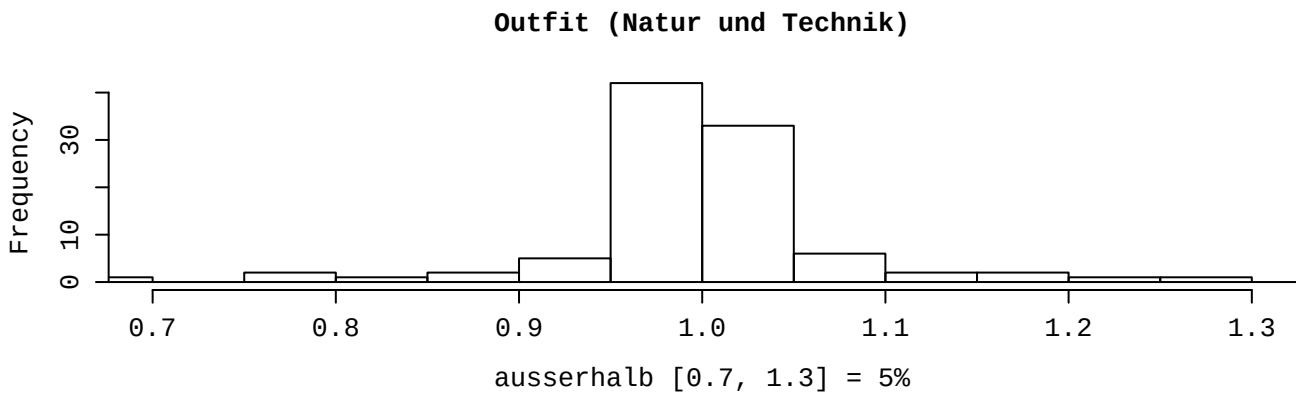

Infit (Englisch Hören)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 5%

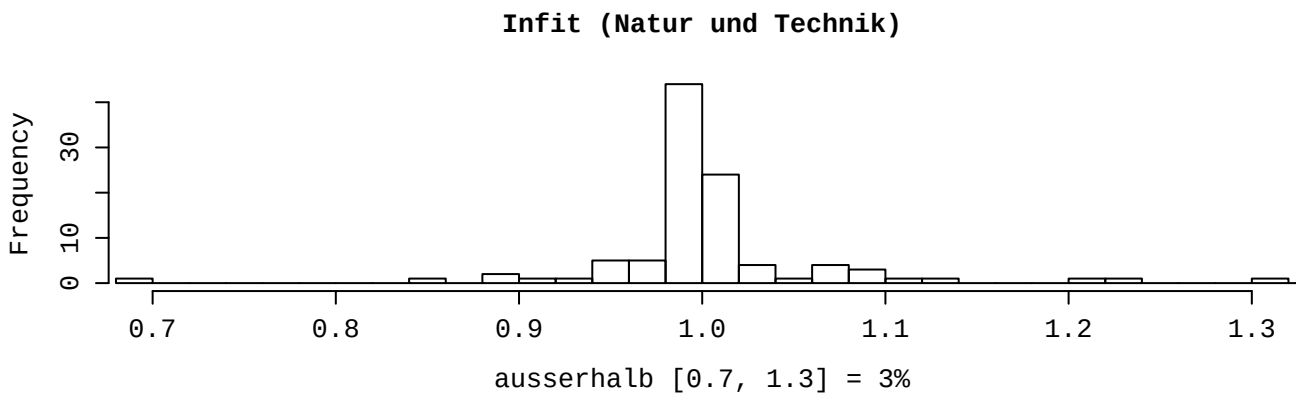

Testinformation pro Hauptpfad (Natur und Technik)
N = 5200
EAP-Rel. = 0.832
1 n=3619
2 n=1128
3 n=277
Information
2
4
6
8
10
12
14
600
700
800
900
1000
1100
1200
Check-Skala


Personenfähigkeit (Natur und Technik)
Frequency
0 50 150
600 700 800 900 1000 1100 1200
Check−Skala M = 819 SD = 47


Itemschwierigkeit (Natur und Technik)
Frequency
0 1 2 3 4 5 6
600 700 800 900 1000 1100 1200
Check−Skala M = 831 SD = 84


Itemtrennschärfe (standardisiert) (Natur und Technik)
Frequency
0 5 10 20
−3 −2 −1 0 1 2 3
M = 0 SD = 1


Outfit (Natur und Technik)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 5%


Infit (Natur und Technik)
Frequency
0
10
30
0.7
0.8
0.9
1.0
1.1
1.2
1.3
ausserhalb [0.7, 1.3] = 3%

Mittelwerte und Standardabweichungen

	dles	dsif	dsch	math	eles	ehoe	fles	fhoe	fsch	natw	espr	fspr
M	841	842	841	844	861	901	813	811	812	816	6	6
SD	63	58	60	55	68	65	64	62	60	47	2	2

M = Mittelwert auf Check-Skala, *SD* = Standardabweichung auf Check-Skala

dles = Deutsch Lesen, *dsif* = Deutsch Sprache im Fokus, *dsch* = Deutsch Schreiben, *math* = Mathematik, *eles* = Englisch Lesen, *ehoe* = Englisch Hören, *fles* = Französisch Lesen, *fhoe* = Französisch Hören, *fsch* = Französisch Schreiben, *natw* = Natur und Technik, *espr* = Englisch Sprechen,
fspr = Französisch Sprechen

Skalen-Interkorrelationen

	dles	dsif	dsch	math	eles	ehoe	fles	fhoe	fsch	natw	espr	fspr
dles	10809	0.69	0.64	0.62	0.68	0.64	0.65	0.54	0.43	0.64	0.45	0.38
dsif	10741	10751	0.69	0.7	0.64	0.59	0.66	0.58	0.57	0.61	0.6	0.36
dsch	10734	10675	10887	0.58	0.54	0.5	0.56	0.48	0.49	0.54	0.29	0.41
math	10737	10684	10764	10849	0.58	0.55	0.59	0.5	0.46	0.71	0.43	0.46
eles	10418	10370	10427	10445	10504	0.81	0.68	0.57	0.5	0.56	0.66	0.27
ehoe	10425	10377	10434	10454	10489	10514	0.63	0.56	0.46	0.54	0.63	0.27
fles	9048	9021	9068	9071	8960	8956	9117	0.71	0.62	0.57	0.56	0.47
fhoe	9081	9052	9100	9103	8994	8993	9107	9150	0.62	0.49	0.26	0.6
fsch	9037	9009	9113	9063	8953	8950	8996	9027	9136	0.38	0.41	0.56
natw	5852	5825	5854	5871	5763	5765	5351	5373	5330	5910	0.38	0.19
espr	124	124	124	124	120	122	84	84	84	39	124	0.34
fspr	91	89	91	90	82	81	89	86	91	46	51	91

Oberhalb Diagonale: Korrelationen nach Pearson (r)
Unterhalb Diagonale: Anzahl Fälle pro Variablenpaar

dles = Deutsch Lesen, *dsif* = Deutsch Sprache im Fokus, *dsch* = Deutsch Schreiben, *math* = Mathematik, *eles* = Englisch Lesen, *ehoe* = Englisch Hören, *fles* = Französisch Lesen, *fhoe* = Französisch Hören, *fsch* = Französisch Schreiben, *natw* = Natur und Technik, *espr* = Englisch Sprechen,
fspr = Französisch Sprechen