

swissuniversities

swissuniversities

Effingerstrasse 15, Postfach

3001 Bern

www.swissuniversities.ch

Inscription et admission aux études de médecine

Mise en œuvre de la procédure de sélection 2023

**Rapport cadre de swissuniversities à l'attention de la
Conférence suisse des hautes écoles**

Table des matières

1.	Mandat et remerciements	3
2.	Capacité d'accueil et inscription aux études de médecine	3
	2.1. Capacité d'accueil	3
	2.2. Inscriptions 15 février	3
3.	Participation au test d'aptitudes pour les études en médecine 2022	4
	3.1. Participation au test totale et par discipline	4
	3.2. Participation au test par sexe et discipline	5
	3.3. Participation par canton	6
4.	Déroulement du test / résultats du test	7
5.	Procédure d'attribution des places d'études	7
	5.1. Admissions par discipline	7
	5.2. Réorientations entre lieux d'études	8
	5.3. Retraits après attribution des places	8

Annexe

Annexe 1 : Rapport scientifique du Centre pour le développement de tests et le diagnostic de l'Université de Fribourg concernant la mise en œuvre et les résultats du test AMS 2023

Abréviations

AMS	Test d'aptitudes pour les études en médecine
CP	Chiropractie
CSHE	Conférence suisse des hautes écoles
CTD	Centre pour le développement de tests et le diagnostic de l'Université de Fribourg
NC	Numerus Clausus

1. Mandat et remerciements

Le 25 novembre 2021, le Conseil des hautes écoles de la Conférence suisse des hautes écoles (CSHE) a approuvé les principes et les valeurs de référence de la procédure d'admission aux études de médecine. Il a chargé swissuniversities de lui soumettre un rapport annuel sur la mise en œuvre de la procédure de sélection. L'élément central du rapport de swissuniversities est le rapport scientifique du Centre pour le développement de tests et le diagnostic de l'Université de Fribourg (CTD) sur la mise en œuvre du test d'aptitudes AMS.

Le nombre d'inscriptions, les places d'études disponibles, la répartition par sexe, par langue et par discipline sont présentés dans le rapport cadre de swissuniversities. Le rapport du CTD, en revanche, se concentre sur le développement, la mise en œuvre et l'évaluation du test.

swissuniversities remercie toutes les personnes impliquées dans la réalisation du test, en particulier les collaborateurs du CTD et des hautes écoles ayant participé au projet, pour leur engagement et leur précieuse collaboration.

2. Capacité d'accueil et inscription aux études de médecine

Le nombre de places d'études en médecine (médecine humaine y compris chiropractie, médecine dentaire et vétérinaire) est limité, notamment en raison du coût élevé des études de médecine et du nombre restreint de places de formation pratique dans les hôpitaux et les cliniques. Depuis des années, le nombre d'inscriptions aux études de médecine dans les hautes écoles suisses dépasse considérablement les capacités d'accueil.

La sélection des étudiants est effectuée de différentes manières : Dans les universités de Genève, Lausanne et Neuchâtel, la sélection a lieu après la première année d'études par le biais d'un « concours ». Les universités de Bâle, Berne, Fribourg, Zurich (y compris les filières de Lucerne et St. Gall), l'Università della Svizzera italiana et l'Ecole polytechnique fédérale de Zurich limitent l'accès aux études de médecine au moyen du numerus clausus (NC) et du test d'aptitudes aux études en médecine (AMS).

2.1. Capacité d'accueil

Au semestre d'automne 2023, 15 places supplémentaires en médecine humaine étaient disponibles en comparaison à l'année précédente au niveau bachelor à l'Università della Svizzera italiana en coopération avec l'Université de Berne ; les capacités sont restées inchangées en médecine dentaire pour 2023/2024 et une augmentation de six places en médecine vétérinaire a été enregistrée au niveau bachelor à l'Université de Berne.

En médecine humaine, 2 187 places d'études étaient donc disponibles dans toute la Suisse en 1^{ère} année de bachelor et 1 460 en 1^{ère} année de master, contre respectivement 172 et 136 places en médecine vétérinaire et 172 et 140 en médecine dentaire.

2.2. Inscriptions 15 février

Jusqu'à l'échéance du 15 février 2023, 7 204 inscriptions au total ont été reçues pour toutes les hautes écoles et toutes les disciplines. Il y a donc eu un peu plus de préinscriptions que l'année précédente. En 2022, le nombre d'inscriptions était plus élevé que ce qui était attendu, et en 2022, il se situait à nouveau dans la tendance à long terme. Le nombre d'inscriptions était considérablement plus élevé que la capacité dans toutes les disciplines.

La figure 1 montre l'évolution des inscriptions dans les hautes écoles sans numerus clausus (universités de Genève, Lausanne et Neuchâtel) sur 10 ans.



Figure 1: Inscriptions dans les hautes écoles sans NC 2014-2023

En 2023, les inscriptions dans les hautes écoles avec numerus clausus (universités de Bâle, Berne, Fribourg, Zurich, Università della Svizzera italiana et EPFZ) représentaient, toutes disciplines confondues, 331 % des capacités disponibles (5 165 inscriptions contre 1 561 places d'études).

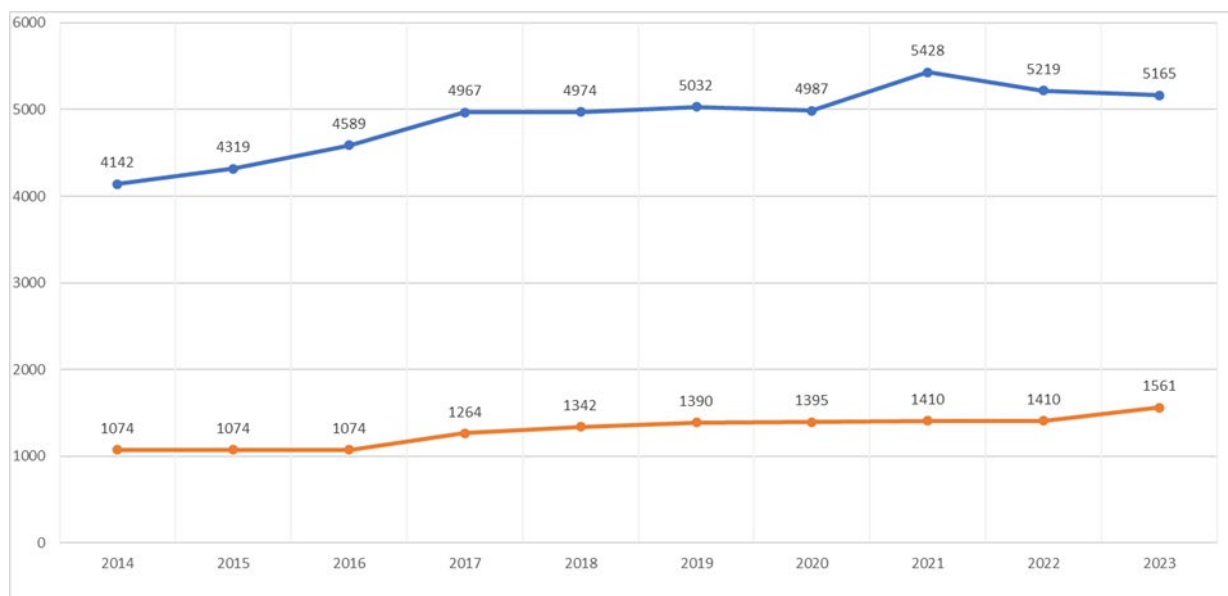


Figure 2: Inscriptions (bleu) et capacités (rouge) dans les hautes écoles avec NC 2014-2023

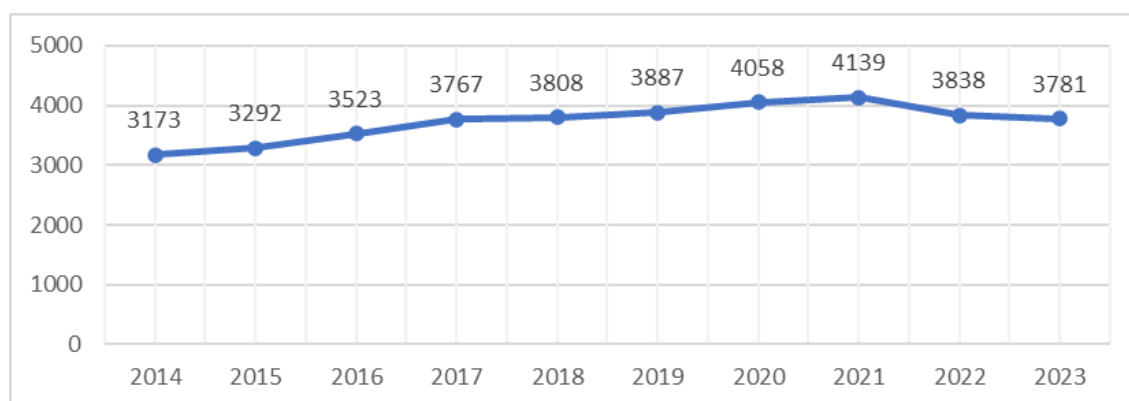
3. Participation au test d'aptitudes pour les études en médecine 2022

Après le délai d'inscription du 15 février, il y a chaque année des désistements. Certaines des personnes inscrites ne s'immatriculent pas (hautes écoles sans NC) ou ne s'inscrivent pas à l'AMS (hautes écoles avec NC). L'expérience montre qu'il y a également d'autres retraits après l'inscription à l'AMS. En 2023, 5 165 personnes se sont inscrites jusqu'au 15 février dans une haute école avec NC, 3 904 d'entre elles se sont inscrites à l'AMS jusqu'au délai de fin mai, 3 782 ont participé au test, 3 781 ont finalement eu un résultat de test valable.

L'AMS peut être passé en allemand, en italien et en français. Comme les universités de Genève, Lausanne et Neuchâtel ne limitent pas l'accès à la première année d'études au moyen d'un NC, les personnes de langue maternelle française (337 personnes, 9% des participants au test) ou italienne (139 personnes, 4%) sont moins nombreuses, relativement à la population résidente, à participer au test que celles de langue maternelle allemande (3 305 personnes, 87%). La médecine vétérinaire n'est proposée en Suisse qu'à Zurich et Berne, c'est pourquoi la proportion de francophones est plus élevée dans cette discipline (124 personnes, 30%).

3.1. Participation au test totale et par discipline

En 2023, le nombre total de participants à l'AMS (toutes disciplines confondues) a diminué par rapport à l'année précédente.



4Figure 3: Participation au test totale 2014-2023

En médecine humaine, le nombre de participations au test a pour la première fois diminué en 2023 par rapport à 2022. En médecine vétérinaire, le nombre de participations aux test était légèrement inférieur, en médecine dentaire, il a de nouveau augmenté.

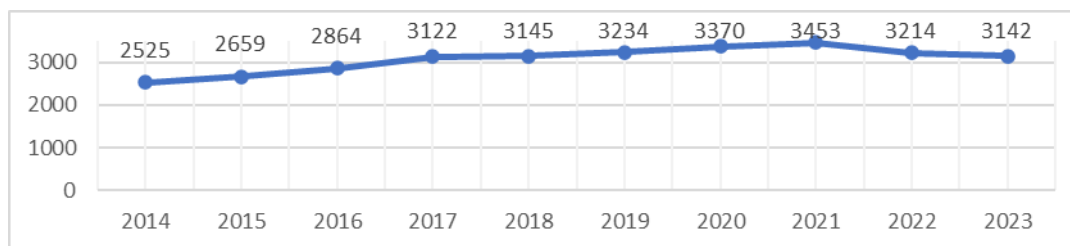


Figure 4: Participation au test médecine humaine y compris chiropractie 2014-2023

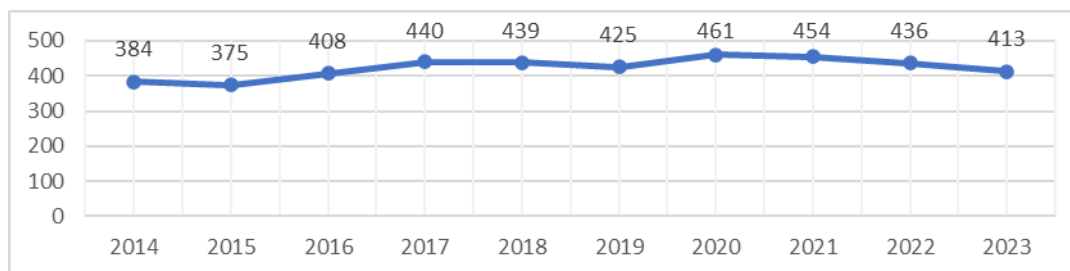


Figure 5: Participation au test médecine vétérinaire 2014-2023

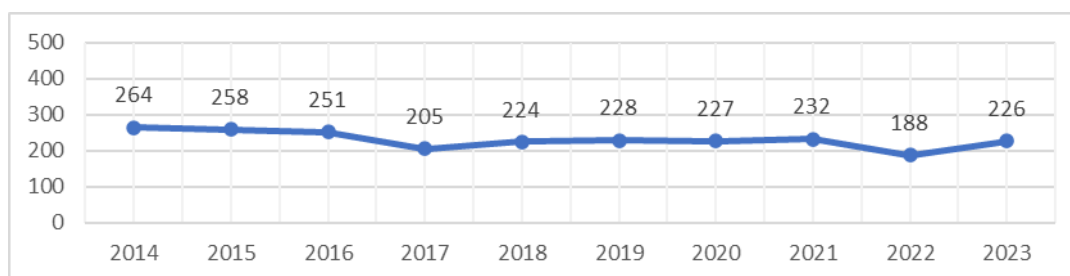


Figure 6: Participation au test médecine dentaire 2014-2023

3.2. Participation au test par sexe et discipline

En 2023, les femmes ont à nouveau été plus nombreuses que les hommes à participer à l'AMS. En médecine humaine (y compris la chiropractie), 2 085 femmes et de 1 057 hommes ont participé au test, en médecine vétérinaire 357 femmes et 56 hommes et en médecine dentaire 165 femmes et 61 hommes.

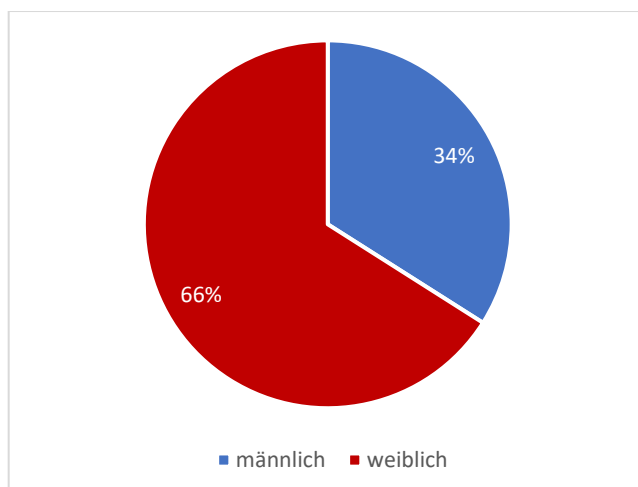


Figure 7: Participation 2023 par sexe en médecine humain (y c. CP)

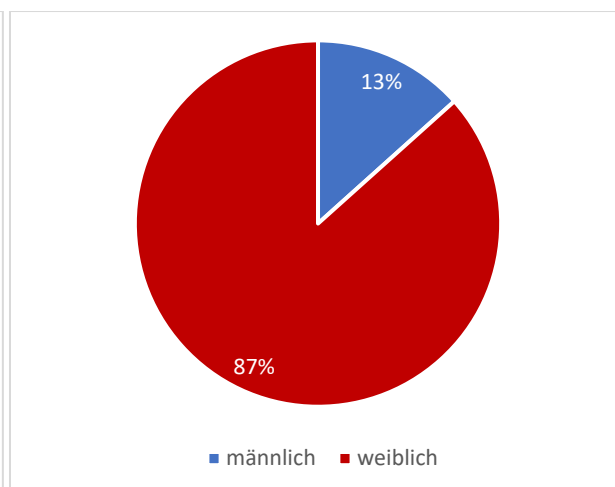


Figure 8: Participation 2023 par sexe en médecine vétérinaire

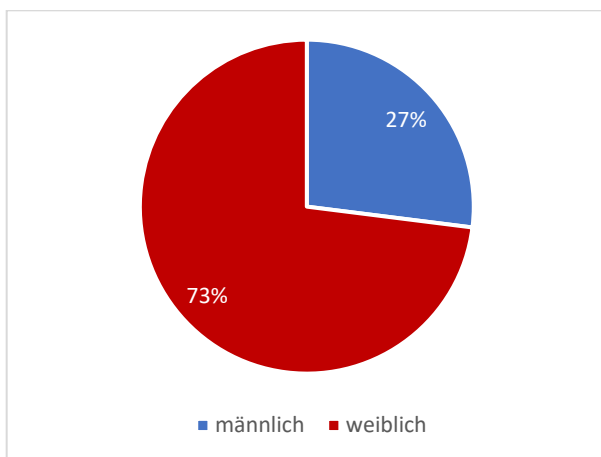


Figure 9: Participation 2023 par sexe en médecine dentaire

3.3. Participation par canton

Parmi les 3 781 personnes qui ont participé à l'AMS en 2023 et ont eu un résultat de test valable, 59 résidaient à l'étranger et 3 722 en Suisse. Le tableau ci-dessous présente la répartition par canton (domicile).

Canton	Participant·e·s
AG	338
AI	5
AR	24
BE	515
BL	222
BS	153
FR	154
GE	31
GL	11
GR	92
JU	23
LU	245
NE	19
NW	15
OW	17
SG	259
SH	31
SO	138
SZ	78
TG	105
TI	131
UR	7
VD	80
VS	89
ZG	98
ZH	842
Etranger	59
Total	3 781

Tableau 1: Participation 2023 par canton

4. Déroulement du test / résultats du test

En 2023, l'AMS avait pour la deuxième fois une nouvelle structure avec 18 tâches par groupe de tâches. La nouvelle structure est plus longue que la version réalisée durant la pandémie, mais plus courte que la version originale du test. Cette réduction permet de maintenir la réalisation du test sans interruption à midi.

Le rapport sur le déroulement et les résultats du test est présenté dans le rapport scientifique du Centre pour le développement de tests et le diagnostic de l'Université de Fribourg (CTD) sur la réalisation de l'AMS 2023, qui se trouve en annexe du présent rapport cadre de swissuniversities.

5. Procédure d'attribution des places d'études

Après l'AMS, swissuniversities procède à l'attribution des places d'études en se basant sur les résultats du test calculés par le CTD (rang percentile et rang moyen). L'attribution se fait automatiquement dans l'outil d'inscription MEDON. Dans un premier temps, le plus grand nombre possible de candidats et candidates se voit attribuer une place d'étude dans chaque discipline, de sorte que la capacité d'admission totale de toutes les hautes écoles soit épuisée. En plus des participants au test de l'année en cours, il faut également tenir compte des candidats qui ont fait valoir leur résultat au test de l'année précédente (2023 : 101 candidats·e·s).

5.1. Admissions par discipline

En médecine humaine (y compris la chiropractie), 61 % des candidats et candidates aux études n'ont pas obtenu de place (1 966 personnes). En médecine vétérinaire, 56 % des candidatures (233 personnes) n'ont pas pu être admises, en médecine dentaire 46 % (115 personnes).

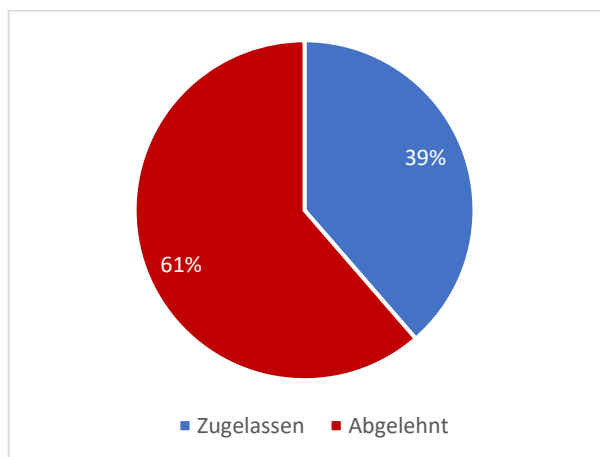


Figure 10: Admissions 2023 médecine humaine (y c. chiropractie)

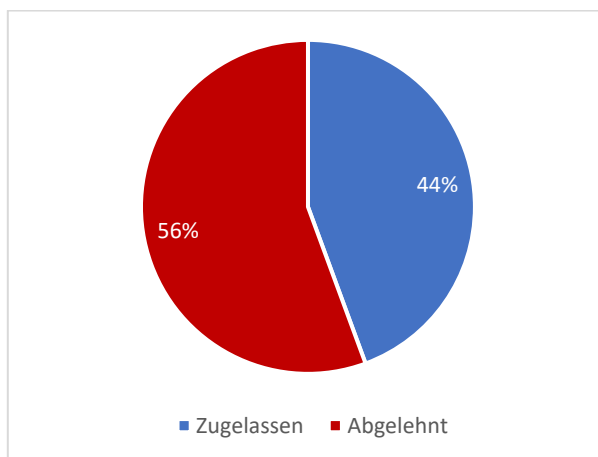


Figure 11: Admissions 2023 médecine vétérinaire

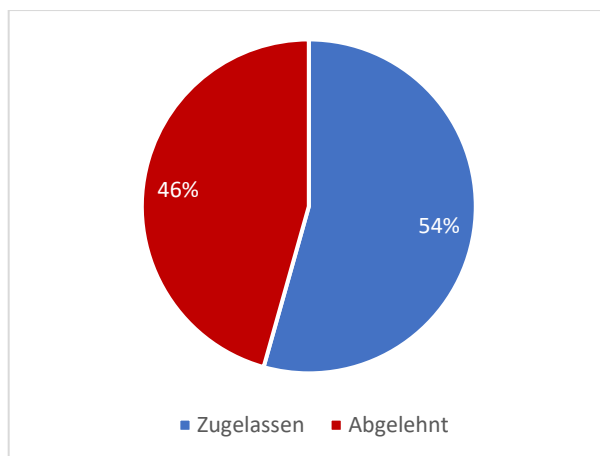


Figure 12: Admissions 2023 médecine dentaire

5.2. Réorientations entre lieux d'études

Dans une deuxième étape, la répartition entre les lieux d'études est également effectuée dans l'outil d'inscription MEDON. Dans la mesure du possible, swissuniversities tient compte des souhaits des candidats et de candidates.

En médecine humaine, la diversification de l'offre d'études et la demande variée entraînent, en 2023 également, de nombreuses réorientations entre les hautes écoles. Plusieurs hautes écoles sont à la fois émettrices et réceptrices de personnes réorientées. En médecine dentaire, la proportion de réorientations est également relativement élevée. La médecine vétérinaire ne peut être étudiée que dans deux lieux d'études, c'est pourquoi la part de réorientations est nettement plus faible.

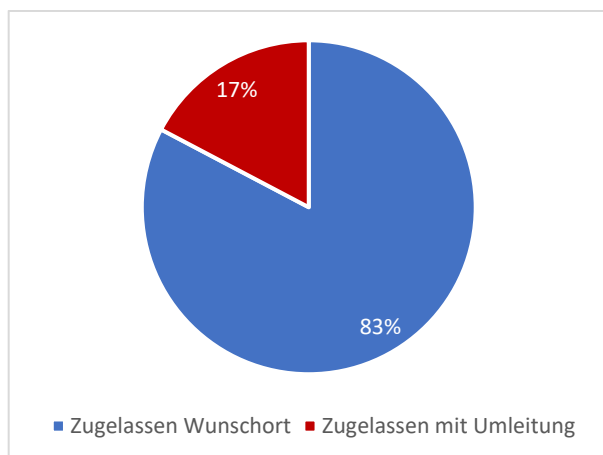


Figure 13: Réorientations 2023 médecine humaine)

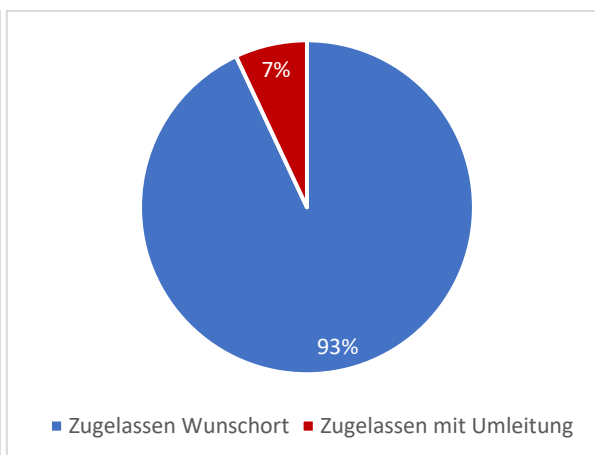


Figure 14: Réorientations 2023 médecine vétérinaire

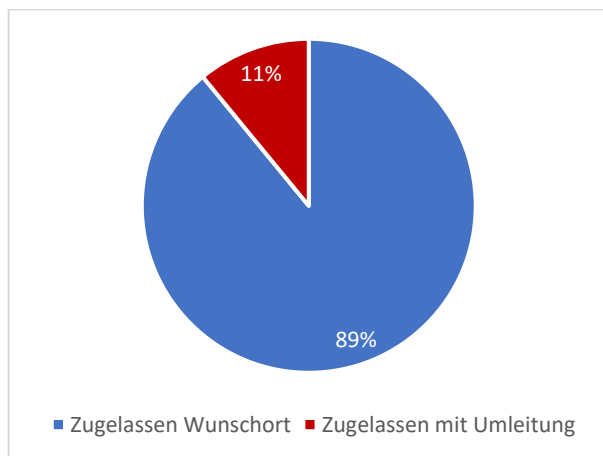


Figure 15: Réorientations 2023 médecine dentaire

Le pourcentage de candidats et de candidates qui ont obtenu une place, mais pas dans la haute école de leur choix, est relativement stable depuis 2017 et se situe globalement (toutes disciplines confondues) entre 15 et 20 %.

5.3. Retraits après attribution des places

En 2023, jusqu'à dix jours avant le début des cours, 129 candidats et de candidates ont renoncé à la place qui leur avait été attribuée, dont 110 en médecine humaine, neuf en médecine vétérinaire et dix en médecine dentaire.

swissuniversities a interrogé ces 129 personnes à propos de leurs motifs. 79 ont répondu à l'enquête. Il s'avère qu'une année de pause est la raison la plus fréquente pour un renoncement (51 mentions), y compris 7 personnes qui accomplissent le service militaire ou civil et 4 personnes qui veulent d'abord terminer d'autres études. 14 personnes ont opté pour d'autres études, 10 n'étaient pas satisfaites du lieu de la réorientation et 5 n'ont pas pu commencer leurs études pour des raisons de santé ou motifs personnels survenus à court terme. Seuls 5 d'entre eux souhaitent étudier à l'étranger et 4 voulaient savoir s'ils avaient de bons résultats au test, sans réel projet d'études. 45 personnes souhaitent que le résultat du test soit reporté à l'année prochaine. Seules 14 personnes ne le feront certainement pas et 20 n'ont pas encore pris de décision.

2023 / BERICHT 30

EIGNUNGSTEST FÜR DAS MEDIZINSTUDIUM (EMS)

BERICHT ÜBER DIE DURCHFÜHRUNG UND ERGEBNISSE

Dr. B. Spicher

Dr. Spicher, Benjamin (2023)
EMS Eignungstest für das Medizinstudium 2023
Berichte des Zentrums für Testentwicklung, Nr. 30,
im Auftrag der Schweizerischen Hochschulkonferenz
(SHK)

Unter Mitarbeit von M. Bernasconi, T. Cruchaud,
L. Equey, M. Strazzeri und Y. de Zordo

In Zusammenarbeit mit der Rektorenkonferenz der
Schweizer Hochschulen (swissuniversities), die das
Anmelde- und Zuteilungsverfahren der Plätze zum
Medizinstudium durchführt und für diesen Bericht
statistische Angaben zur Verfügung stellt:
K. Balmer

Redaktion: T. Cruchaud



1 Inhalt

1	Beschreibung des verwendeten Eignungstests.....	4
1.1	Berechnung der Werte.....	6
1.2	Mittlerer Rangplatz der Aufgabengruppen.....	7
2	Testanwendung 2023	9
2.1	Verteilungsprüfung.....	9
2.2	Äquivalenz der Sprachversionen	10
2.2.1	<i>Sprachvergleich für die Aufgabengruppen</i>	<i>12</i>
2.2.2	<i>Darstellung des Korrekturverfahrens.....</i>	<i>16</i>
2.2.3	<i>Identifikation von DIF-Aufgaben und Bestimmung des Korrekturwertes.....</i>	<i>18</i>
2.3	Effekte der Korrektur.....	24
2.4	Vergleichbarkeit der Testlokale	27
2.5	Vergleich für die Geschlechter.....	29
2.6	Vergleiche nach Wunschhochschulen.....	31
3	Ergebnisse zur Testgüte.....	32
3.1	Zuverlässigkeit.....	32
3.2	Binnenstruktur.....	33
4	Literatur.....	35
4.2	Frühere Berichte des ZTD	36

1 Beschreibung des verwendeten Eignungstests

Beim EMS handelt es sich um eine Adaptation des deutschen **Tests für Medizinische Studiengänge (TMS)** mit Weiterentwicklungen (durch das ZTD) für die Schweiz. Der TMS mit 9 Aufgabengruppen hat sich in Deutschland bewährt und wird dort seit 1986 eingesetzt (1986 bis 1996 deutschlandweit, seit 2007 erneut für ausgewählte Universitäten). In der Schweiz wird der Test seit 1998 erfolgreich angewendet – bis 2003 mit identischer Struktur wie beim deutschen TMS, seit 2004 durchlief der EMS verschiedene Anpassungen.

Seit 2022, nach zwei Durchführungen als spezielle, „coronabedingte“ Versionen, besteht der EMS wiederum aus neun Aufgabengruppen (*Fakten lernen* und *Figuren lernen* sind jeweils in eine Einpräg- und eine Reproduktionsphase aufgeteilt), die zu einem Gesamtwert verrechnet werden. Ablauf und Struktur des aktuellen Tests sind in Tabelle 1 dargestellt.

Tabelle 1: Struktur und Ablauf des EMS 2023

EMS Struktur (seit 2022)	Max. Punkte	Zeit in Min	Zeit pro Aufgabe (min)
Aufgabengruppe			
Muster zuordnen	18	16	0,9
Medizinisch-naturwissenschaftliches Grundverständnis	18	45	2,5
Schlauchfiguren	18	10	0,6
Quantitative und formale Probleme	18	45	2,5
Figuren einprägen		4	
Fakten einprägen		6	
Wechsel von Testheft A zu Testheft B			
Textverständnis	18	45	2,5
Figuren reproduzieren	18	5	0,3
Fakten reproduzieren	18	6	0,3
Diagramme und Tabellen	18	45	2,5
Konzentriertes und sorgfältiges Arbeiten	18	8	
Total	162	235	

In jeder Aufgabengruppe können maximal 18 Punkte erzielt werden. Die Aufgabengruppen *Muster zuordnen*, *Fakten lernen*, *Figuren lernen* sowie *Konzentriertes und sorgfältiges Arbeiten* werden jährlich durch das Zentrum für Testentwicklung neu erstellt, die übrigen Aufgaben werden in Deutschland einmalig vorerprobt und für den ersten zulassungsrelevanten Einsatz im EMS verwendet.

Jedes Jahr werden neue Aufgaben für alle Aufgabengruppen entwickelt. Die Anforderungen an die Aufgaben sind hoch, u.a. (1) müssen sie jedes Jahr die Studieneignung gleich zuverlässig messen, (2) muss das Schwierigkeitsspektrum aller Aufgaben annähernd vergleichbar sein, (3) darf kein spezielles Fachwissen vorausgesetzt werden, um die Trainierbarkeit des Tests gering zu halten und (4) muss eine eindeutige, richtige Lösung existieren.

Ausgangspunkt der Aufgabenentwicklung des TMS/EMS war eine differenzierte Anforderungsanalyse des Medizinstudiums, an der Lehrbeauftragte sowie Experten und Expertinnen mitarbeiteten. Das Ziel war die Vorhersage des Studienerfolges, um ein faires und wissenschaftlich begründetes Zulassungsverfahren zu erhalten. Auf der Grundlage dieser Analysen wurden ca. 50 Bereiche und Aufgabentypen geprüft und die unter vielen Aspekten am besten geeignet erscheinenden 13 für die Erprobungen im Übergangsverfahren des TMS ausgewählt. Von diesen haben vor allem aufgrund der Ergebnisse zur Validität dann 9 Aufgabengruppen Eingang in den TMS und den EMS gefunden.

Die Stärken des EMS lassen sich folgendermassen zusammenfassen:

- Die Auswahl der Testanforderungen erfolgt aus einer grösseren Menge möglicher Studienanforderungen aufgrund einer Anforderungsanalyse. Dabei werden ständige Anpassungen an sich verändernde Bedingungen geprüft und wo nötig vorgenommen.
- Der wissenschaftliche Nachweis der Vorhersagbarkeit des Studienerfolgs mittels EMS ist gewährleistet. Dies erlaubt die gesetzlich geforderte Studienzulassung nach Eignung.
- Die Konstruktion der Aufgaben erfolgt durch Experten und es findet eine anschliessende empirische Überprüfung statt, damit die Aufgaben alle Gütekriterien erfüllen, die gewünschte Fähigkeit tatsächlich messen und optimal „schwierig“ sind.
- Für die Beantwortung der Aufgaben ist kein spezielles fachliches Vorwissen notwendig, es wird tatsächlich die „Studierfähigkeit“ als aktuelle Fähigkeit zur Wissensaneignung und Problemlösung gemessen. Dadurch ist der Test auch wenig trainierbar, was sich positiv auf die Sozialverträglichkeit auswirkt (kein zusätzlicher Gewinn durch zusätzliche Trainingskurse nachgewiesen, wenn die empfohlene Vorbereitung erfolgt).
- Es ist ein ökonomisches und genau kapazitätsentsprechendes Zulassungsverfahren möglich, die Rechtsfähigkeit des Verfahrens hat sich bei mehreren Überprüfungen (auch gerichtlich) bestätigt.

1.1 Berechnung der Werte

Alle Aufgabengruppen des EMS liefern Punkte als Summe richtig gelöster Aufgaben zwischen 0 und 18. Es können somit maximal 162 Punkte erreicht werden. Basierend auf den Kennwertanalysen wurden 2023 alle Aufgaben gewertet.

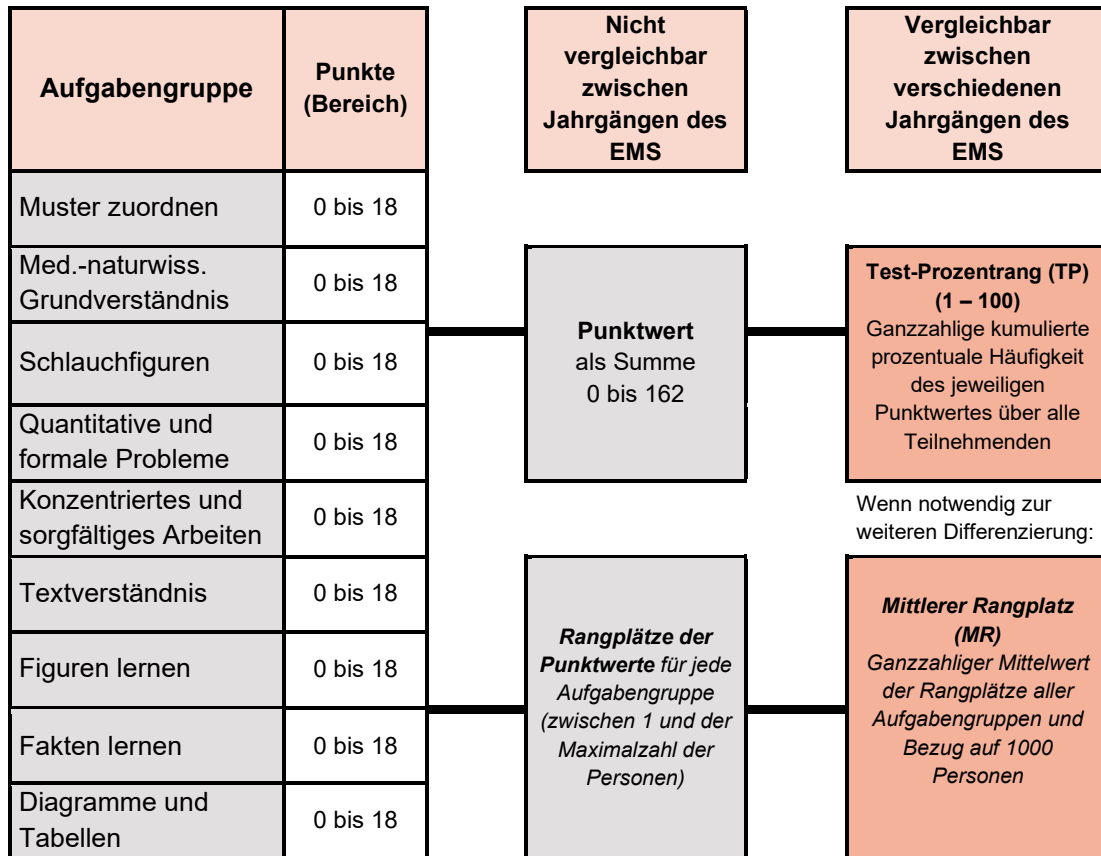


Abbildung 1: Punktwerte der einzelnen Aufgabengruppen 2023 und ihre Zusammenführung über den Punktwert zum Test-Prozentrang und mittleren Rangplatz.

Die Aufgabengruppe *Konzentriertes und sorgfältiges Arbeiten* besteht aus 1600 Zeichen, wovon 400 nach alljährlich wechselnden Regeln als Zielzeichen zu markieren sind. Es können allerdings in der zur Verfügung stehenden Zeit kaum alle Zeichen bearbeitet werden. Die Position des letzten markierten Zeichens bestimmt, wie viele Zeichen als bearbeitet gewertet werden. Alle übersehenen und fälschlich markierten Zeichen vor diesem letzten bearbeiteten Zeichen zählen als Fehler und werden von der Menge der korrekt markierten Zeichen abgezogen. Der resultierende Wert ist der „Rohwert“, der in eine Skala zwischen 0 und 18 transformiert wird, um mit den anderen Aufgabengruppen gleichgewichtig zum Punktwert addiert zu werden. Die schlechtesten 1.5% (aber bis zu mindestens Rohwert 0) erhalten den Punktwert 0, die besten 1.5% erhalten den Punktwert 18. Der Wertebereich des Rohwertes für die verbleibenden 97% wird in 17 äquidistante Abschnitte eingeteilt und den Punktwerten 1 bis 17 zugeordnet. Obwohl sich die Testanforderung zwischen den Jahren ändert, resultieren so vergleichbare Punktwertverteilungen.

Alle Punkte der Aufgabengruppen werden zu einer Summe addiert (Punktwert). Dieser Wert hat den Nachteil, dass er aufgrund schwankender Schwierigkeiten nicht zwischen Tests verschiedener Jahre vergleichbar ist.

Aus der Gesamtpunktzahl (Punktwert) werden deshalb zwei Werte berechnet:

- **der Test-Prozentrang (TP) und**
- **der mittlere Rangplatz der Aufgabengruppen (MR)**

Der Test-Prozentrang (TP) zeigt an, wie viele Prozent aller teilnehmenden Personen ein schlechteres oder gleich gutes Ergebnis erreicht haben. Die Differenz zu 100 zeigt an, wie viele Prozent der Personen ein besseres Testergebnis erreicht haben. Er wird direkt aus der kumulierten Häufigkeits-Verteilung der Punktwerte aller Personen berechnet, die in einem Jahr am EMS-Test teilnahmen und ist ganzzahlig. Werte unter 10 werden auf 10 gerundet. Test-Prozentrang und mittlerer Rangplatz sind über die Jahre vergleichbar – dies ist Voraussetzung dafür, dass ein Testergebnis ins Folgejahr übernommen werden kann.

1.2 Mittlerer Rangplatz der Aufgabengruppen

Nur bei gleichem Test-Prozentrang werden die Personen in der Reihenfolge des mittleren Rangplatzes aller Aufgabengruppen (MR) berücksichtigt, wenn nicht mehr alle mit entsprechendem Test-Prozentrang zugelassen werden können. Diese Rangplätze werden über alle Aufgabengruppen gemittelt. Niedrige mittlere Rangplätze entsprechen jeweils den besseren Leistungen. Der Rangplatz ist umso niedriger, je höher die Punktzahl ist und je weniger Personen insgesamt einen besseren Punktwert erreicht haben. Rangplätze belohnen gute Leistungen in einer schwierigen Aufgabengruppe stärker, indem dort niedrigere Rangplätze für die besten Leistungen resultieren. Zur Herstellung der Vergleichbarkeit der Rangplätze zwischen den Jahren wird der mittlere Rangplatz auf die Basis von 1000 Personen umgerechnet.

Eine Beispielrechnung zur Ermittlung des mittleren Rangplatzes wird nachfolgend dargestellt.

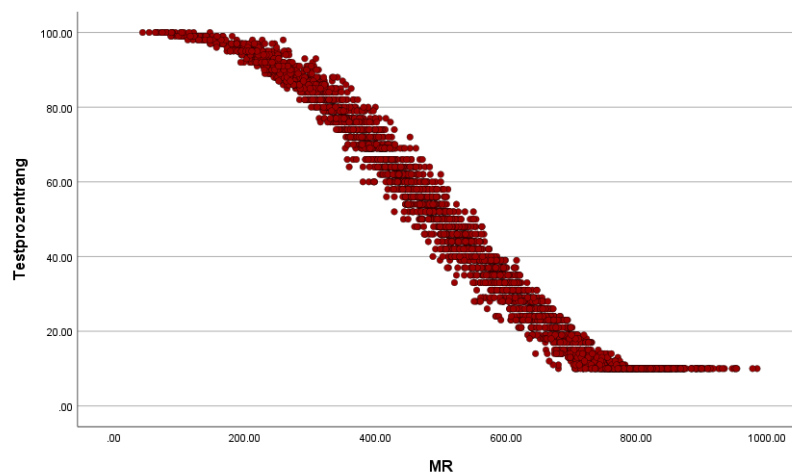


Abbildung 2: Beziehung zwischen Test-Prozentrang (TP) und mittlerem Rangplatz (MR) für 2023 (Werte kleiner 10 auf TP 10 angehoben).

Der mittlere Rangplatz MR bewegt sich 2023 zwischen 44 und 984. Niedrige Werte stehen für bessere Leistungen. Mittlerer Rangplatz und Test-Prozentrang korrelieren mit -0.98 (trotz der Anhebung aller TP-Werte unter 10 auf 10).

Beispielrechnung

100 Personen nehmen an einem Test mit 2 Aufgabengruppen teil, in denen man je maximal 20 Punkte erreichen kann.

Aufgabengruppe 1:

- Hier erzielen 10 Kandidaten und Kandidatinnen die maximale Punktzahl 20. Bei gleicher Punktzahl erhält jede Person den durchschnittlichen Rang dieser Personen (die 1 bis 10 belegen).
- Kandidatin A hat diese maximale Punktzahl erreicht und belegt damit in dieser Aufgabengruppe zusammen mit den anderen 9 Kandidaten und Kandidatinnen Rang 5.5 (Durchschnitt der Ränge 1 bis 10, also $(1+10)/2=5.5$).
- Kandidat B hat, zusammen mit 14 weiteren Kandidaten und Kandidatinnen, 19 Punkte erzielt. Diese Kandidaten und Kandidatinnen erhalten in dieser Aufgabengruppe den Rang 18 (Durchschnitt der Ränge 11 bis 25, also $(11+25)/2=18$).

Aufgabengruppe 2:

- In dieser Aufgabengruppe erreicht nur Kandidat B die volle Punktzahl von 20 Punkten. Er erhält als einziger den Rang 1.
- Kandidatin A erreicht hier 19 Punkte, zusammen mit nur 3 weiteren Kandidaten und Kandidatinnen. Sie alle erhalten in dieser Aufgabengruppe den durchschnittlichen Rang $(2+5)/2=3.5$.

Kandidatin A und Kandidat B erreichten beide damit die gleiche Punktzahl von 39 Punkten (und den gleichen Test-Prozentrang).

Der mittlere Rangplatz für Kandidatin A und Kandidat B berechnet sich wie folgt:

- Kandidatin A: $(5.5 + 3.5)/2=4.5$
- Kandidat B: $(18 + 1)/2=9.5$

Und auf 1000 Personen:

- Kandidatin A: $4.5/100 \times 1000=45$
- Kandidat B: $9.5/100 \times 1000=95$

Hätten 1000 Personen am Test teilgenommen, resultiert für Kandidatin A im Mittel über beide Aufgabengruppen der 45. Rang, für Kandidat B der 95. Rang. Kandidatin A wird dafür belohnt, in der ersten Aufgabengruppe die maximale, aber vor allem auch in der zweiten Aufgabengruppe zusammen mit nur sehr wenigen Kandidaten und Kandidatinnen eine hohe Punktzahl erreicht zu haben.

Kandidat B erhält durch die geringere Punktzahl in der ersten Aufgabengruppe (in dem viele Personen das gleiche oder höhere Ergebnis erzielt haben) einen höheren (schlechteren) Rang.

Der Unterschied zwischen Kandidatin A und Kandidat B ist dadurch gerechtfertigt, dass Kandidatin A trotz gleicher Punktzahl über beide Aufgabengruppen mehr Kandidaten und Kandidatinnen übertroffen hat als Kandidat B.

Müsste man zwischen beiden Personen unterscheiden, wäre Kandidatin A die bessere.

2 Testanwendung 2023

2.1 Verteilungsprüfung

Die Verteilung des Punktwerts entspricht annähernd einer Normalverteilung. Metrische Prüfverfahren sind bedingt aussagefähig. Die Punktwerte der einzelnen Aufgabengruppen sind statistisch nicht als normalverteilt anzusehen (Tabelle 2).

Tabelle 2: Kennwerte der Punktwertskalen Gesamt und Aufgabengruppen

	Punktwert	Quantitative und formale Probleme	Schlauchfiguren	Textverständnis	Med.-naturwiss. Grundv.	Figuren lernen	Fakten lernen	Muster zuordnen	Diagramme und Tabellen	Konzentr. und sorgf. Arbeiten
Mittelwert	81.1	8.5	11.8	7.0	8.4	10.6	9.8	8.9	6.6	9.4
Median	81.0	8.0	12.0	7.0	8.0	11.0	10.0	9.0	6.0	10.0
Modalwert	85.0	8.0	13.0	6.0	9.0	11.0	9.0	8.0	6.0	11.0
Stand.-abweichung	18.6	3.0	3.5	2.8	2.7	3.4	3.6	3.2	2.8	4.1
Spannweite	123.0	18.0	17.0	17.0	16.0	18.0	18.0	18.0	18.0	18.0
25. Perzentil	68.0	6.0	9.0	5.0	7.0	8.0	7.0	7.0	5.0	7.0
75. Perzentil	94.0	11.0	14.0	9.0	10.0	13.0	12.0	11.0	8.0	12.0
Schiefe	-0.1	0.1	-0.5	0.4	0.0	-0.1	0.2	0.2	0.4	-0.3
Exzess	-0.3	-0.1	-0.4	-0.1	-0.2	-0.4	-0.6	-0.1	-0.1	-0.2

Für die Aufgabengruppe *Diagramme und Tabellen* wurden 2023 die tiefsten mittleren Punktwerte registriert, gefolgt von *Textverständnis*. Die höchsten mittleren Punktwerte lieferten *Schlauchfiguren* und *Figuren lernen*. Alle Aufgabengruppen differenzieren gut bis sehr gut im relevanten Wertebereich.

2.2 Äquivalenz der Sprachversionen

Sprachrelevante Aufgaben wurden nach den gleichen Kriterien wie in den Vorjahren von der deutschen Vorlage adaptiert. Eine ausführliche Beschreibung des Verfahrens kann dem Bericht 5 (1999) des ZTD entnommen werden. Zielsetzung bleibt, möglichst übereinstimmende Aufgabenschwierigkeiten (vergleichbare Ergebnisse bei vergleichbaren Fähigkeiten) zwischen den Sprachversionen zu schaffen.

Die Chancengleichheit als Gleichbehandlung der Sprachgruppen muss gewährleistet sein, die Lösungswahrscheinlichkeit einer Aufgabe muss bei gleicher Fähigkeit tatsächlich auch vergleichbar sein. Die Schlussredaktionen der Testversionen in französischer und italienischer Sprache werden von mehreren nativ französisch- bzw. italienischsprachigen Personen durchgeführt. Diese wurden dahingehend geschult, dass es vor allem auf eine Übertragung der Schwierigkeit in die jeweilige Sprache ankommt.

Die Gleichheit ist nicht allein durch Übersetzung zu beeinflussen, sondern hängt auch von kulturellen und weiteren Faktoren ab. Bei allen Adaptationen muss mit Unterschieden hinsichtlich der Aufgabenschwierigkeiten zwischen den Sprachversionen gerechnet werden.

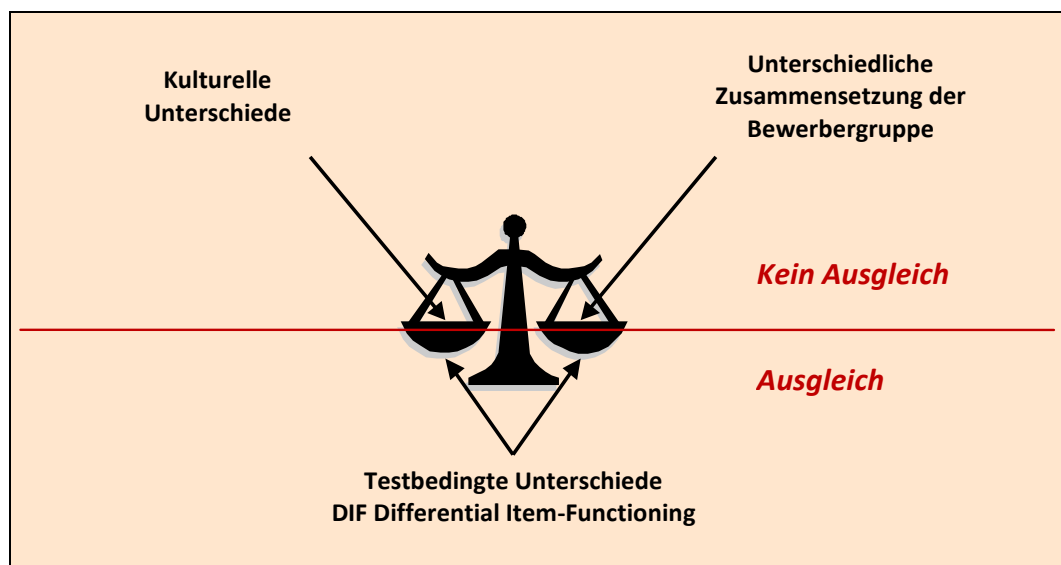


Abbildung 3: Faktoren mit Wirkung auf Unterschiede zwischen den Sprachversionen.

Zwei Ursachen (kulturelle Unterschiede und die unterschiedliche Zusammensetzung der Bewerbergruppe) führen potenziell zu „echten“ Personen-Unterschieden, die nicht auf den Test oder seine Adaptation zurückgeführt werden können. Sie müssen durch die Ergebnisse entsprechend abgebildet werden – hier erfolgt kein Ausgleich.

Testbedingte Unterschiede hingegen können durch die Adaptation verursacht sein und sollen möglichst vollständig ausgeglichen werden. Mögliche Ursache wäre die Vereinfachung oder Erschwerung der Aufgabe bei der Übersetzung (Wortwahl, Satzgliederung o.ä.). Erneut wird zu diesem Zweck das bewährte DIF-Verfahren (Differential Item-Functioning) angewendet.

Kulturelle Unterschiede zwischen den Sprachgruppen

- Mögliche generelle Fähigkeitsunterschiede oder unterschiedliche Fähigkeitsprofile (anderes Verhältnis der einzelnen Fähigkeiten zueinander) im Vergleich der Kulturen können vorhanden sein, die möglicherweise auf Unterschiede im Schulsystem zurückzuführen sind. Auch die Strategien, wie entsprechende Aufgaben gelöst werden, oder die Art und Intensität der Vorbereitung können sich unterscheiden und auf die Ergebnisse auswirken.
- Unterschiede in Leistungsvoraussetzungen (z.B. Motivation, Belastbarkeit, Ausdauer) können zu unterschiedlichen Resultaten beitragen.

Unterschiedliche Repräsentativität der Stichproben für die Sprachgruppen

- Die jeweils untersuchten Stichproben können verschiedene Ausschnitte aus der jeweiligen Sprachgruppe repräsentieren. Französisch- und italienischsprachige Personen können den NC durch ein Studium in ihrer Muttersprache an einer anderen Universität vermeiden. Die geringe Zahl der französisch- und italienischsprachigen Teilnehmenden macht dies wahrscheinlich: Die Auswahl kann die „Spitze“ oder das „Ende“ der Leistungsrangreihe aller Maturanden und Maturandinnen der jeweiligen Sprachgruppe überrepräsentieren, etwa durch unterschiedliches Wahlverhalten aufgrund der vorhandenen Alternativen für Studienorte.

Testbedingte Unterschiede

- Die Testaufgaben können nach der Übertragung eine unterschiedliche Aufgabenschwierigkeit aufweisen, indem durch Satzstellung, Wortwahl, Kompliziertheit des Satzes etc. ein Unterschied auftritt. Die Lösungsschwierigkeit einer Aufgabe wird unter anderem auch von der Formulierung und dem Satzbau einer Fragestellung beeinflusst. Bereits geringe Änderungen innerhalb einer Sprache können zu unterschiedlichen Schwierigkeiten führen. Allerdings sind diese Differenzen nicht vorherzusehen, sondern können erst empirisch nachgewiesen werden.

Der Beschluss der Schweizerischen Hochschulkonferenz aus dem Jahre 1999 ist weiterhin bindend. Er beinhaltet:

1. Einen Ausgleich nur bei sprachabhängigen Tests vorzunehmen. Dies sind fünf von neun Aufgabengruppen. Der Sprachausgleich beschränkt sich demnach auf die stärker sprachabhängigen Aufgabengruppen *Medizinisch-naturwissenschaftliches Grundverständnis, Quantitative und formale Probleme, Textverständnis, Fakten lernen* sowie *Diagramme und Tabellen*.
2. Den Ausgleich nur vorzunehmen, wenn signifikante Mittelwertunterschiede in der entsprechenden Aufgabengruppe vorhanden sind. Geprüft werden dazu die Abweichungen der jeweiligen Zielsprache von der deutschen Sprachgruppe.

Adaptationsbedingte Effekte sind nur in den sprachabhängigen Aufgabengruppen zu erwarten, während die sprachunabhängigen Aufgabengruppen davon kaum betroffen sein können.

Spricht man von DIF, so muss dieses von systematischen Mittelwertsdifferenzen, welche alle Aufgaben einer Gruppe gleichermaßen betreffen, abgegrenzt werden. Die Ergebnisse der letzten Jahre haben gezeigt, dass sich die Sprachgruppen in manchen Aufgabengruppen signifikant unterscheiden. Differenzen kommen nicht nur durch einzelne Aufgaben zustande, sondern auch durch eine Verschiebung der Schwierigkeiten aller Aufgaben einer Gruppe. Es ist unwahrscheinlich, dass diese systematischen Unterschiede testbedingt sind. Testbedingte Unterschiede als sprachliche Besonderheiten zeigen sich in spezifischen Abweichungen einzelner Aufgaben. Systematische Unterschiede sind dagegen mit grösserer

Wahrscheinlichkeit den Faktoren „Kulturunterschiede“ und „unterschiedliche Repräsentativität“ zuzurechnen.

Es ergibt sich für das Ausgleichsverfahren die nachfolgende Zielstellung:

Das Risiko, dass testbedingte Unterschiede die Sprachunterschiede systematisch beeinflussen, soll verringert werden. Andererseits soll kein Ausgleich von Unterschieden aufgrund unterschiedlicher Zusammensetzung der Stichproben und der kulturellen Unterschiede erfolgen.

2.2.1 Sprachvergleich für die Aufgabengruppen

Auch 2023 treten Unterschiede zwischen den Sprachgruppen in sprachabhängigen und wenig sprachabhängigen Aufgabengruppen auf – die grössten Unterschiede sind bei der sprachfreien Aufgabengruppe *Schlauchfiguren* und der sprachabhängigen Aufgabengruppe *Diagramme und Tabellen* zwischen der deutschen und französischen Sprachgruppe zu beobachten. Die unkorrigierten Punktzahlen nach Sprachen sind nachfolgend (Abbildung 4) abgebildet.

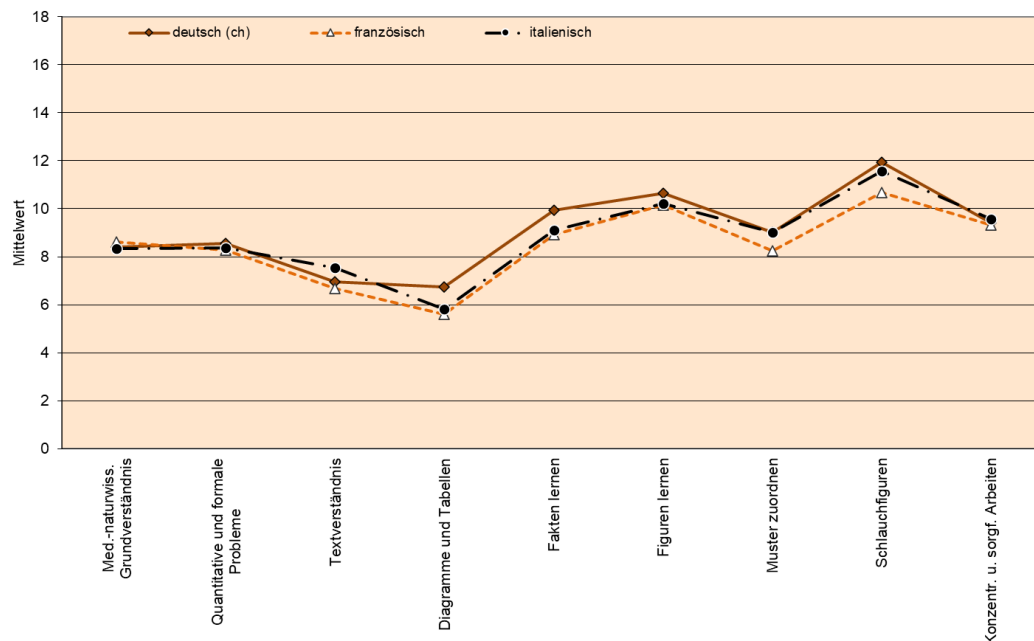


Abbildung 4: Mittelwerte der Punktwerte für die Aufgabengruppen nach Sprachen (unkorrigiert).

Abbildung 5 zeigt den Vergleich zwischen deutsch- und französischsprachigen Testteilnehmenden für das Jahr 2023, in Abbildung 6 ist der Vergleich zwischen deutsch- und italienischsprachigen Personen dargestellt. Die senkrechten Linien bezeichnen Mittelwert und Standardabweichung der Differenzen über die Jahrgänge für die letzten 10 Jahre. Positive Differenzwerte deuten auf bessere Ergebnisse der deutschen Sprachgruppe hin, negative auf solche der jeweils anderen Sprachgruppe.

Für die französische Sprachgruppe treten mehrheitlich deutlich geringere Unterschiede auf, als dies in den letzten 10 Jahren der Fall war.

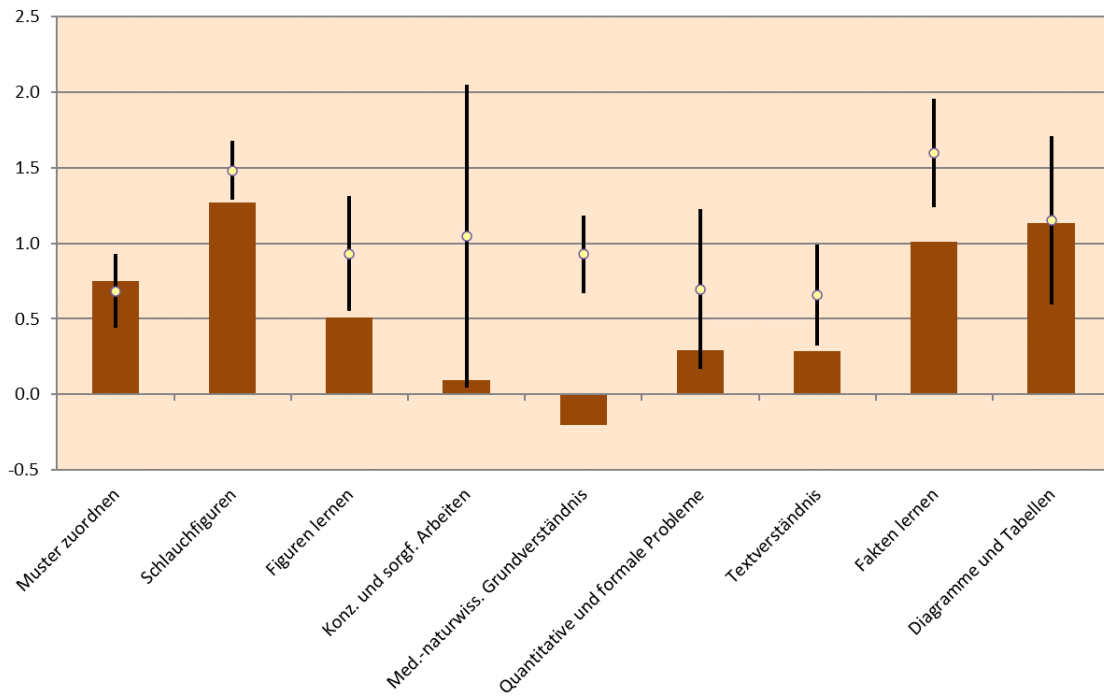


Abbildung 5: Differenz für Aufgabengruppen deutschsprachig-französischsprachig 2023 (unkorrigiert) und Mittelwerte mit Standardabweichungen der Testdurchführungen 2013-2022.

Die Ergebnisse der italienischsprachigen Teilnehmenden liegen mehrheitlich im Rahmen der bekannten Abweichungen. Höhere Abweichungen treten bei *Muster zuordnen*, *Konzentriertes und sorgfältiges Arbeiten* sowie *Medizinisch-naturwissenschaftliches Grundverständnis* auf.

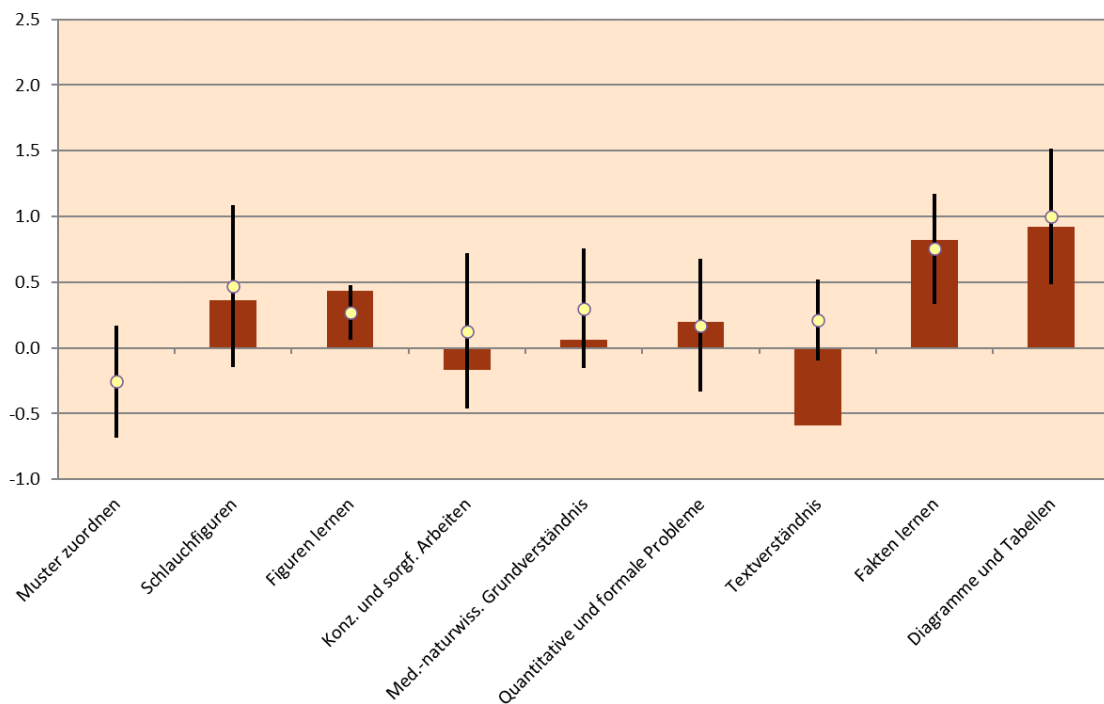


Abbildung 6: Differenz für Aufgabengruppen deutschsprachig-italienischsprachig 2023 (unkorrigiert) und Mittelwerte mit Standardabweichungen der Testdurchführungen 2013-2022.

2.2.1.1 Vergleich von französischsprachigen Personen aus dem Kanton Freiburg und ausserhalb

Bereits im Vorfeld des EMS 2022 kam es zu politischen Diskussionen, die sich um die Zukunft der (französischsprachigen) ärztlichen Versorgung im Kanton Freiburg drehten. Vorgebracht wurde, dass französischsprachige Personen durch die verfügbaren Vorbereitungsangebote in deutscher Sprache beim EMS benachteiligt wären. Swissuniversities und das ZTD propagieren andererseits seit Jahren die Durchführung von Probelaufen in allen drei Testsprachen und stellen entsprechende Unterlagen allen Schulen zur Verfügung. Frühere Evaluationen der Vorbereitungsstrategien der Teilnehmenden haben gezeigt, dass das Angebot solcher Probelaufe dennoch insbesondere in französischsprachigen Regionen weniger bekannt ist und genutzt wird. Diesem Umstand wurde, neben den bereits bestehenden kostenfreien Angeboten, 2022 zusätzlich entgegengewirkt, indem das bereits bestehende Angebot an Probelaufen im Kanton Freiburg stärker propagiert wurde. Zusätzlich wurde an der Universität Freiburg ein weiterer Probelauf durch „NC-Wiki“ angeboten. Dabei handelt es sich um eine studentische, nicht gewinnorientierte Initiative. 2023 wurden diese Angebote weiter ausgebaut und propagiert.

Sollten die durch swissuniversities und ZTD unterstützten Massnahmen Wirkung zeigen, müsste sich dies insbesondere auf die im Kanton Freiburg wohnhaften Personen in einem grösseren Ausmass auswirken, da zu erwarten wäre, dass der Anteil teilnehmender Personen aus diesem Kanton gegenüber anderen Kantonen erhöht war.

In den Vorjahren unterschieden sich diese Personen nicht signifikant von „ausserkantonale“ wohnhaften, französischsprachigen KandidatInnen. Nachfolgend (Abbildung 7) sind die entsprechenden Werteverteilungen der unkorrigierten Punktwerte für Humanmedizin und Veterinärmedizin (zu geringe Teilnehmerzahlen für Zahnmedizin) für die Sprachgruppen deutsch „ohne Freiburg“, deutsch „Freiburg“, französisch „Freiburg“ und französisch „ohne Freiburg“ dargestellt. Zwischen der deutschen Kohorte und den französischsprachigen KandidatInnen aus dem Kanton Freiburg treten wie bereits 2022 auch 2023 in beiden Disziplinen keine signifikanten Mittelwertsdifferenzen auf, wohl aber zwischen den deutschsprachigen Personen und den französischsprachigen Testteilnehmenden, die nicht im Kanton Freiburg wohnhaft sind.

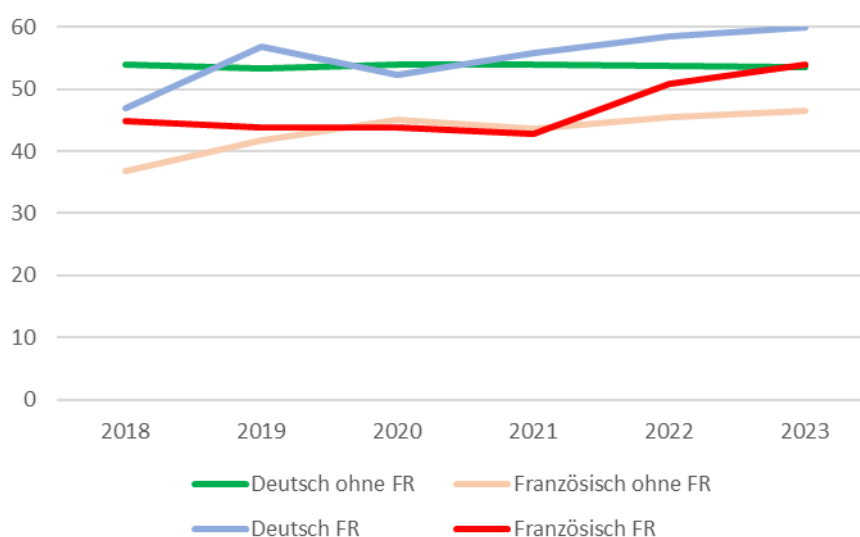


Abbildung 7: Entwicklung der mittleren Testprozenträge für deutsche und französische Sprachgruppen über die letzten 6 Jahre, Humanmedizin

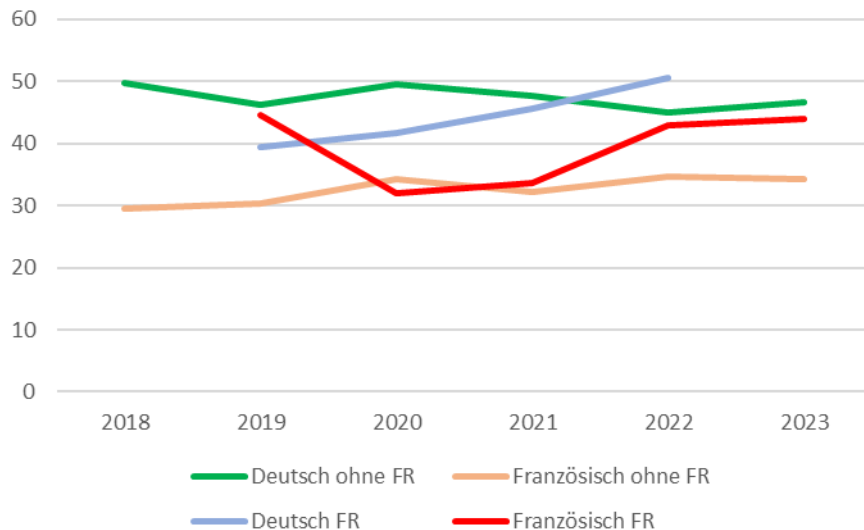


Abbildung 8: Entwicklung der mittleren Testprozenträge für deutsche und französische Sprachgruppen über die letzten 6 Jahre, Veterinärmedizin

Die Differenzen zwischen den beiden französischsprachigen Gruppen erreichen aufgrund der geringen Gruppengrösse der im Kanton Freiburg wohnhaften Personen nur auf Ebene einiger Aufgabengruppen signifikantes Niveau, die erreichten Punktwerte liegen jedoch praktisch durchwegs über denjenigen der französischsprachigen Personen aus anderen Kantonen.

Für die mindestens tendenziell unterschiedlichen Leistungen der beiden Gruppen können verschiedene Ursachen relevant sein. Der Umstand, dass zwischen den Ergebnissen von deutschsprachigen und im Kanton Freiburg wohnhaften französischsprachigen Personen in diesem Jahr keine signifikanten Unterschiede bezüglich der erreichten Punktwerte (mehr) auftreten, spricht hingegen zumindest nicht gegen die Wirksamkeit von Probeläufen – und relativiert gleichzeitig die oft propagierte hohe Effizienz kostenpflichtiger Vorbereitungsangebote, da letztere ja in französischer Sprache nach wie vor kaum verfügbar sind.

Tabelle 3 weist die detaillierten Werte und Mittelwertsvergleiche für die einzelnen Aufgabengruppen aus. Zwischen deutschsprachigen und im Kanton Freiburg wohnhaften französischsprachigen Personen treten auch auf Ebene der Aufgabengruppen kaum (mit Ausnahme von *Konzentriertes und sorgfältiges Arbeiten*) signifikante Mittelwertsdifferenzen auf. In drei Aufgabengruppen erreichen die französischsprachigen Personen aus dem Kanton Freiburg signifikant (trotz geringer Gruppengrösse) bessere Ergebnisse, als französischsprachige Personen aus anderen Kantonen.

Tabelle 3: Signifikanzprüfungen für Prozenträge und Aufgabengruppen: d, f(FR), f(a)

Sprachgruppe		Mittelwert	Stand.abw.	Minimum	Maximum
Test-Prozenrang $d^{**} > f(a)$	d	52.0	28.1	10	100
	f (a)	41.1	25.9	10	99
	f (FR)	52.0	28.5	10	99
Muster zuordnen $d^{**} > f(a)$	d	9.0	3.2	0	18
	f (a)	8.1	3.0	2	18
	f (FR)	8.7	3.0	1	15
MNGV	d	8.4	2.7	0	17
	f (a)	8.6	2.6	1	17
	f (FR)	8.7	2.8	3	15
Schlauchfiguren $d^{**} > f(a)$	d	11.9	3.5	1	18
	f (a)	10.3	3.5	2	18
	f (FR)	11.6	3.9	3	18
QfP $d^* > f(a)$	d	8.6	3.0	0	18
	f (a)	8.0	3.2	1	18
	f (FR)	9.0	2.8	4	16
Textverständnis $d^{**} > f(a)$	d	7.0	2.8	0	17
	f (a)	6.5	2.5	1	15
	f (FR)	7.3	2.5	2	14
Figuren lernen $d^{**} > f(a)$	d	10.7	3.4	0	18
	f (a)	9.9	3.4	2	18
	f (FR)	10.8	3.4	5	18
Fakten lernen $d^{**} > f(a)$	d	9.9	3.7	0	18
	f (a)	8.8	3.1	3	18
	f (FR)	9.4	3.4	3	18
DuT $d^{**} > f(a)$	d	6.7	2.8	0	18
	f (a)	5.4	2.4	14	14
	f (FR)	6.2	2.3	1	12
Konz. und sorgf. Arbeiten	d	9.4	4.1	0	18
	f (a)	9.3	3.8	0	18
	f (FR)	9.4	3.8	0	18

n: d: deutsch 3343 / f(FR): französisch Freiburg 58 / f(a): französisch andere 289

**/*: signifikanter Einfluss des Faktors „Sprachgruppe“ bei varianzanalytischer Prüfung (1%- bzw. 5%-Niveau); angegeben sind signifikante Unterschiede bei multiplen Mittelwertvergleich.

Die Effizienz des Probelauf-Angebots dürfte nicht alleine auf den eigentlichen Anlass zurückzuführen sein. Vielmehr ist anzunehmen, dass bereits das Vorhandensein entsprechender Angebote in der geografischen Nähe die Bereitschaft zur intensiveren Vorbereitung erhöht. Auch können Personen, die selber nicht teilnehmen, beispielsweise über Freunde und Bekannte, die vom Probelauf erzählen, zu grösseren Vorbereitungsanstrengungen motiviert werden. Die genauen Wirkmechanismen dürften vielschichtig sein – in jedem Fall scheinen sie zu funktionieren.

2.2.2 Darstellung des Korrekturverfahrens

Das nachfolgend dargestellte Korrekturverfahren bleibt weiterhin für alle nicht-deutschsprachigen Testabsolvierungen relevant, da sichergestellt werden muss, dass durch die sprachliche Adaptation des Tests keinerlei Benachteiligungen auftreten.

2.2.2.1 Identifikation und Ausgleich testbedingter Unterschiede geschehen wie folgt:

- Sprachausgleiche werden nur in den sprachabhängigen Aufgabengruppen *Medizinisch-naturwissenschaftliches Grundverständnis, Quantitative und formale*

Probleme, Textverständnis, Fakten lernen sowie *Diagramme und Tabellen* vorgenommen. Dass Unterschiede in den nichtsprachlichen Tests testbedingt sind (und ausgeglichen werden müssen), ist unwahrscheinlich. Die Aufgaben selber erfordern praktisch keine Sprachkompetenzen und die sprachspezifischen Anforderungen der Instruktionen scheinen so einfach, dass Verständnisunterschiede unwahrscheinlich sind.

- Damit ein Sprachausgleich in einer bestimmten sprachabhängigen Aufgabengruppe erfolgt, müssen sich die entsprechenden mittleren Punktzahlen zwischen den Sprachgruppen signifikant unterscheiden (Vergleich deutsch/französisch, beziehungsweise deutsch/italienisch mittels t-Test). Wenn keine Mittelwert- und Varianzunterschiede vorhanden sind, trägt die Aufgabengruppe auch nicht zu Mittelwertunterschieden hinsichtlich des Test-Prozentrangs bei. Eine Korrektur allfälliger DIF-Aufgaben (die vorhanden sein können) würde dann eventuell zusätzliche Unterschiede produzieren.
- Es werden jene Aufgaben ausgeglichen, welche über eine systematische Differenz hinaus DIF aufweisen, die also eine andere Position in der Rangreihe der Itemschwierigkeiten einnehmen. Für diese Aufgaben ist der Verdacht am ehesten begründet, dass es sich um testbedingte Differenzen handelt. Der Ausgleich wird nicht symmetrisch vorgenommen, da es darum geht, eine Benachteiligung der Vergleichsgruppe gegenüber der Referenzgruppe zu vermeiden.
- Als Sprachausgleich wird den französisch- und italienischsprachigen Probanden die Differenz zur entsprechenden Aufgabenschwierigkeit in der deutschsprachigen Version gutgeschrieben, falls sie die Aufgabe nicht korrekt gelöst haben. Um jede Benachteiligung auszuschliessen, wird nicht nur die Differenz zur Regressionsgeraden ausgeglichen. Die Lage dieser Regressionsgeraden wird auch von allen DIF-Aufgaben mitbestimmt. Man kann das so interpretieren, dass für DIF-Aufgaben bei falscher Antwort zumindest der Schwierigkeitsunterschied zur Referenzgruppe ausgeglichen wird.

Eine generelle Gutschrift des Mittelwertunterschiedes pro Aufgabengruppe für die französisch- und italienischsprachige Gruppe könnte für einzelne Personen zu mehr Punkten pro Aufgabengruppe führen, als theoretisch maximal vergeben werden, und sich auf die Gewichtung der Aufgabengruppen negativ auswirken. Aus diesem Grund wird lediglich die Differenz der Aufgabenschwierigkeit gutgeschrieben. Nur Personen, welche die betreffende Aufgabe nicht korrekt gelöst haben, erhalten diesen Bonus.

2.2.2.2 Welche Aufgabengruppen ausgleichen?

Die folgenden Tabellen zeigen die Sprachvergleiche für die sprachabhängigen Aufgabengruppen. Referenz ist die deutschsprachige Testversion. Die Gesamtheit der französischsprachigen Teilnehmenden erzielte in vier der sprachabhängigen Aufgabengruppen signifikant geringere Punktwerte als die deutschsprachigen Teilnehmenden. Für die italienische Sprachgruppe sind zwei Aufgabengruppen betroffen.

Tabelle 4: Signifikanzprüfung deutsche vs. französische Sprachform (sprachabhängige Aufgabengruppen).

Aufgabengruppe	Sprache	n	m	s	Sig.	DIF-Prüfung
Med.-naturwiss. Grundverständnis	d	3305	8.40	2.732	f > d	Nein
	f	337	8.61	2.640		
Quantitative und form. Probleme	d	3305	8.56	2.960	.045	Ja
	f	337	8.27	3.106		
Textverständnis	d	3305	6.96	2.817	.027	Ja
	f	337	6.68	2.534		
Fakten lernen	d	3305	9.93	3.689	.000	Ja
	f	337	8.92	3.161		
Diagramme und Tabellen	d	3305	6.73	2.806	.000	Ja
	f	337	5.60	2.431		

Tabelle 5: Signifikanzprüfung deutsche vs. italienische Sprachform (sprachabhängige Aufgabengruppen).

Aufgabengruppe	Sprache	n	m	s	Sig.	DIF-Prüfung
Med.-naturwiss. Grundverständnis	d	3305	8.40	2.732	.391	Nein
	i	139	8.34	2.709		
Quantitative und form. Probleme	d	3305	8.56	2.960	.221	Nein
	i	139	8.36	2.941		
Textverständnis	d	3305	6.96	2.817	i > d	Nein
	i	139	7.55	2.790		
Fakten lernen	d	3305	9.93	3.689	.005	Ja
	i	139	9.11	3.396		
Diagramme und Tabellen	d	3305	6.73	2.806	.000	Ja
	i	139	5.81	2.629		

Mit der Bestimmung der zu prüfenden Aufgabengruppen ist noch nicht eruiert, ob es sich bei den Differenzen zwischen den Stichproben um mögliche adaptationsbedingte Unterschiede handelt. Zu diesem Zweck ist die nachfolgend dargestellte Analyse der Aufgaben der betreffenden Aufgabengruppen notwendig.

2.2.3 Identifikation von DIF-Aufgaben und Bestimmung des Korrekturwertes

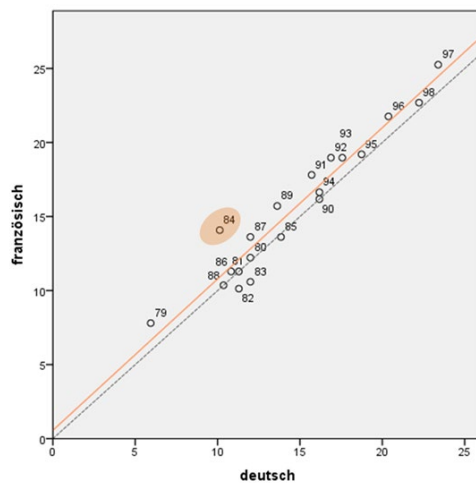
Eine Möglichkeit zur Identifikation von DIF, auch bei relativ kleinen Stichprobengrößen, ist die Methode „Delta-Plot“. Bei diesem Verfahren werden die zu vergleichenden Aufgabenschwierigkeiten z-standardisiert und anschliessend in „Delta-Werte“ transformiert.

Die Transformation erfolgt über die Formel: $\Delta = 13 - 4z$

Dies bedeutet, dass Δ einen Mittelwert von 13 und eine Standardabweichung von 4 aufweist. Hohe Werte stehen für „schwierige“ (von wenigen Probanden und Probandinnen gelöste) Aufgaben.

Die aus den Delta-Werten abgeleitete Regressionsgerade $Y = AX + B$ beschreibt die Beziehung zwischen den analysierten Sprachversionen. Eine graphische Darstellung der Delta-Werte (Delta-Plot) würde im Idealfall eine ellipsenförmige Anordnung von Punkten entlang der Diagonalen ergeben. Dies würde bedeuten, dass sowohl die

Aufgabenschwierigkeiten wie auch deren Reihenfolge in beiden Sprachversionen vergleichbar sind.



Die Diagonale wird in den folgenden Abbildungen durch eine unterbrochene schwarze Linie dargestellt. Ein Abweichen der Werte von dieser Diagonalen ergibt eine Regressionsgerade, welche sich sowohl in Steigung wie Achsenschnittpunkt von der Diagonalen unterscheiden kann. Solche Verschiebungen der Regressionsgeraden stehen für systematische Unterschiede in den Aufgabengruppen, deren Ursachen eher nicht testbedingt (Übersetzung), sondern in realen Gegebenheiten zu vermuten sind. Um die Regressionsgerade gruppierte Aufgaben folgen also diesen Gegebenheiten und sind demzufolge nicht auszugleichen.

Abbildung 9: Beispiel für einen Delta-Plot.

Auffällige Abweichungen einzelner Punkte von dieser Regressionsgeraden hingegen deuten auf Aufgaben hin, welche zusätzlich zu systematischen Verschiebungen spezielle Eigenschaften aufweisen. Bei einer parallelen Verschiebung der Regressionsgeraden bleibt trotz einer Veränderung der absoluten Schwierigkeiten die „Schwierigkeits-Rangfolge“ der Aufgaben erhalten. Von der Regressionsgeraden abweichende Aufgaben stimmen hingegen in der untersuchten Sprachgruppe bezüglich der „Schwierigkeitshierarchie“ nicht mit der Referenzgruppe überein (sie sind im Vergleich zu den anderen Aufgaben zu schwer oder zu leicht ausgefallen). Die Ursache solcher Differenzen kann eher testbedingt, also beispielsweise in der Übersetzung, vermutet werden. Betroffene Aufgaben müssen nicht in jedem Fall als „schlecht“ übersetzt betrachtet werden. Denkbar ist etwa auch, dass in der Fragestellung Konzepte enthalten sein könnten, welche in den Sprachgruppen unterschiedliche Bekanntheitsgrade aufweisen.

Regressionsgeraden werden in der Folge als durchgezogene Linien dargestellt. Massgeblich für die Entscheidung, ob bei einem Item DIF vorliegt, ist die Distanz des entsprechenden Punktes von dieser Geraden. Die Distanz D wird nach der Formel

$$D_i = \frac{AX_i - Y_i + B}{\sqrt{A^2 + 1}}$$

berechnet, wobei unter A die Steigung und B der Achsenschnittpunkt der Regressionsgeraden zu verstehen ist, X_i bezeichnet den Delta-Wert der Referenzgruppe, Y_i denjenigen der zu vergleichenden Gruppe.

In der Folge sollen innerhalb der auszugleichenden Aufgabengruppen die kritischen Aufgaben identifiziert werden. Da, wie oben erwähnt, keine symmetrische DIF-Behandlung verwendet wird, ist die Richtung der Abweichung der Regressionsgeraden von der Diagonalen massgebend. Es werden also jene Aufgaben ausgeglichen, **welche mindestens 1.5 Punkte** (Longford, Holland & Thayer, 1993) von der Regressionsgeraden in Richtung tieferer Punktwerte für die interessierende Sprachgruppe abweichen. Als Sprachausgleich wird den Personen, die die entsprechende Aufgabe nicht gelöst haben, die Differenz der Aufgabenschwierigkeit zur deutschen Aufgabenschwierigkeit gutgeschrieben (also die Differenz der Lösungswahrscheinlichkeit).

2.2.3.1 Quantitative und formale Probleme

Alle Aufgaben liegen eng an der Regressionsgeraden. Trotz signifikanter Differenz hinsichtlich der Punktwerte dieser Aufgabengruppe gibt es auf Aufgabenebene keine Hinweise auf übersetzungsbedingte Einflüsse für die französische Sprachgruppe.

Tabelle 6: DIF-Analyse Quantitative und formale Probleme

Aufgabe	Schwierigkeit			Delta-Wert			Bonus f	Bonus i
	d	f	i	d	f	i		
55	0.82	0.80	-	6.28	6.70	-		-
56	0.73	0.66	-	8.16	9.62	-		-
57	0.64	0.53	-	10.04	12.33	-		-
58	0.76	0.65	-	7.54	9.83	-		-
59	0.66	0.63	-	9.62	10.24	-		-
60	0.57	0.49	-	11.50	13.16	-		-
61	0.70	0.71	-	8.79	8.58	-		-
62	0.51	0.47	-	12.75	13.58	-		-
63	0.44	0.48	-	14.21	13.37	-		-
64	0.51	0.61	-	12.75	10.66	-		-
65	0.37	0.39	-	15.66	15.25	-		-
66	0.32	0.29	-	16.71	17.33	-		-
67	0.34	0.37	-	16.29	15.66	-		-
68	0.26	0.28	-	17.96	17.54	-		-
69	0.21	0.18	-	19.00	19.62	-		-
70	0.16	0.18	-	20.04	19.62	-		-
71	0.28	0.31	-	17.54	16.92	-		-
72	0.29	0.23	-	17.33	18.58	-		-

Aufgabe 64 weist die grösste Differenz hinsichtlich des Delta-Werts auf, allerdings hat die französische Sprachgruppe diese Aufgabe erfolgreicher gelöst als die deutsche.

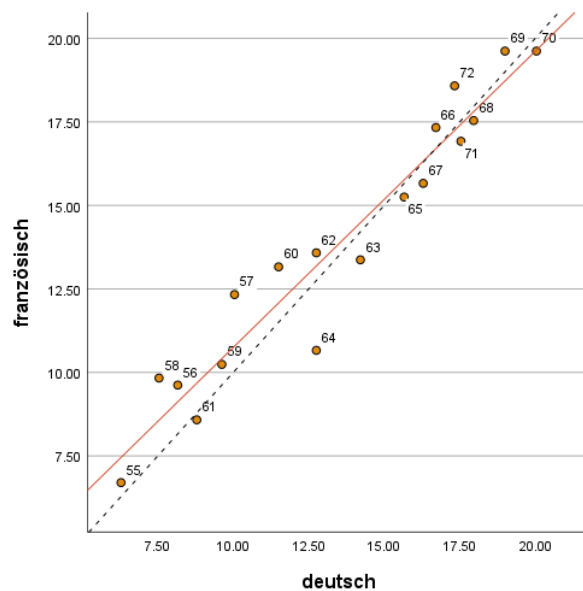


Abbildung 10: Delta-Plot für Quantitative und formale Probleme (deutsch-französisch).

2.2.3.2 Textverständnis

Die Aufgaben liegen eng an der Regressionsgeraden. Einzig Aufgabe 77 weist eine auffällige Differenz zur deutschen Sprachgruppe auf. Den französischsprachigen Personen, die diese Aufgabe nicht korrekt gelöst haben, werden 0.12 Punkte gutgeschrieben.

Tabelle 7: DIF-Analyse Textverständnis

Aufgabe	Schwierigkeit			Delta-Wert			Bonus f	Bonus i
	d	f	i	d	f	i		
73	0.65	0.72		9.83	8.37			-
74	0.47	0.50		13.58	12.95			-
75	0.54	0.45		12.12	14.00			-
76	0.27	0.23		17.75	18.58			-
77	0.46	0.34		13.79	16.29		0.12	-
78	0.23	0.25		18.58	18.17			-
79	0.66	0.70		9.62	8.79			-
80	0.58	0.47		11.29	13.58			-
81	0.52	0.53		12.54	12.33			-
82	0.31	0.39		16.92	15.25			-
83	0.38	0.31		15.46	16.92			-
84	0.23	0.15		18.58	20.25			-
85	0.48	0.53		13.37	12.33			-
86	0.28	0.29		17.54	17.33			-
87	0.20	0.22		19.21	18.79			-
88	0.20	0.11		19.21	21.08			-
89	0.29	0.32		17.33	16.71			-
90	0.20	0.18		19.21	19.62			-

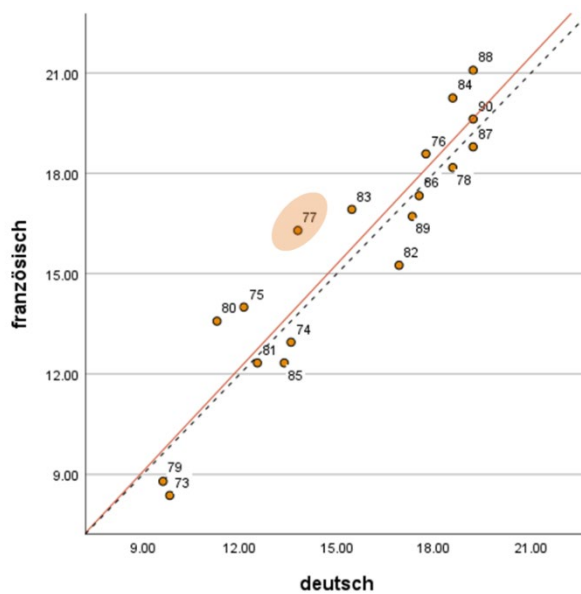


Abbildung 11: Delta-Plot für Textverständnis (deutsch-französisch).

2.2.3.3 Fakten lernen

In der Aufgabengruppe *Fakten lernen* werden der französischen Sprachgruppe für die Aufgabe 114 bei Falschantwort 0.20 Punkte gutgeschrieben. Für die italienischsprachige Gruppe sind es 0.15 Punkte bei einer falsch beantworteten Aufgabe 118.

Tabelle 8: DIF-Analyse *Fakten lernen*

Aufgabe	Schwierigkeit			Delta-Wert			Bonus f	Bonus i
	d	f	i	d	f	i		
109	0.74	0.74	0.72	7.95	7.95	8.37		
110	0.42	0.35	0.34	14.62	16.08	16.29		
111	0.85	0.80	0.76	5.66	6.70	7.54		
112	0.64	0.64	0.49	10.04	10.04	13.16		
113	0.66	0.63	0.67	9.62	10.24	9.41		
114	0.41	0.21	0.32	14.83	19.00	16.71	0.20	
115	0.32	0.29	0.35	16.71	17.33	16.08		
116	0.66	0.65	0.70	9.62	9.83	8.79		
117	0.48	0.46	0.48	13.37	13.79	13.37		
118	0.49	0.40	0.34	13.16	15.04	16.29		0.15
119	0.58	0.50	0.52	11.29	12.95	12.54		
120	0.38	0.31	0.40	15.46	16.92	15.04		
121	0.60	0.53	0.53	10.87	12.33	12.33		
122	0.70	0.69	0.62	8.79	8.99	10.45		
123	0.63	0.55	0.58	10.24	11.91	11.29		
124	0.49	0.42	0.50	13.16	14.62	12.95		
125	0.53	0.41	0.43	12.33	14.83	14.41		
126	0.36	0.34	0.35	15.87	16.29	16.08		

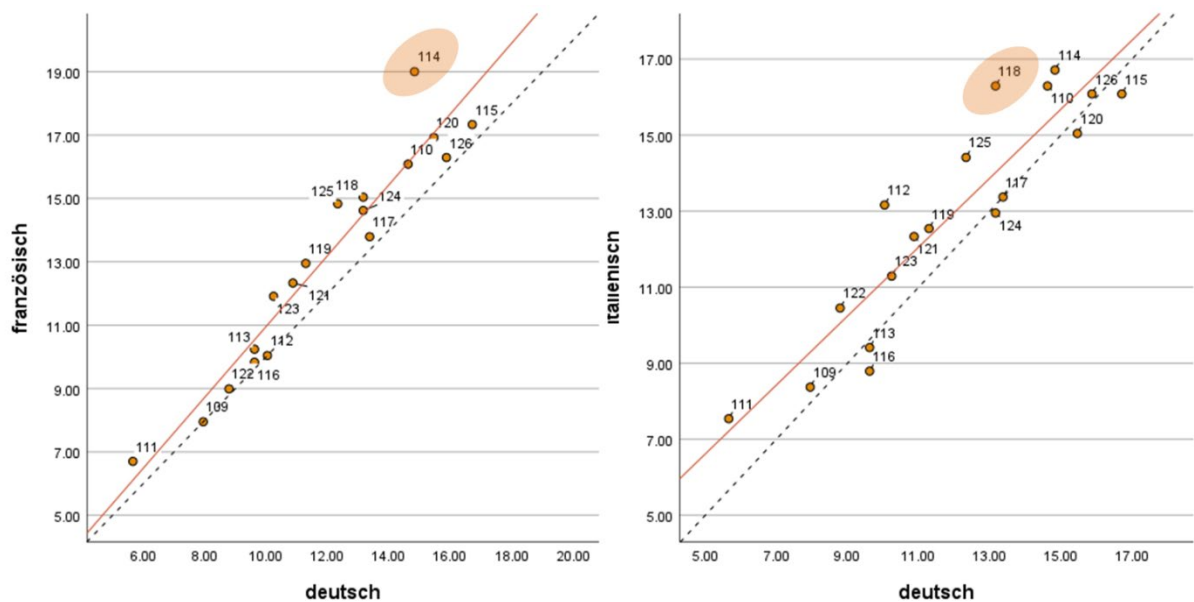


Abbildung 12: Delta-Plot für *Fakten lernen* (deutsch-französisch und deutsch-italienisch).

2.2.3.4 Diagramme und Tabellen

In der Aufgabengruppe *Diagramme und Tabellen* werden der italienischen Sprachgruppe bei Falschantwort für die Aufgabe 128 0.19 Punkte gutgeschrieben. In der französischsprachigen Gruppe fällt keine Aufgabe auf.

Tabelle 9: DIF-Analyse *Diagramme und Tabellen*

Aufgabe	Schwierigkeit			Delta-Wert			Bonus f	Bonus i
	d	f	i	d	f	i		
127	0.81	0.68	0.83	6.49	9.20	6.08	-	
128	0.72	0.57	0.53	8.37	11.50	12.33	-	0.19
129	0.40	0.34	0.40	15.04	16.29	15.04	-	
130	0.46	0.37	0.35	13.79	15.66	16.08	-	
131	0.53	0.45	0.54	12.33	14.00	12.12	-	
132	0.40	0.33	0.36	15.04	16.50	15.87	-	
133	0.45	0.30	0.33	14.00	17.12	16.50	-	
134	0.42	0.26	0.41	14.62	17.96	14.83	-	
135	0.42	0.42	0.47	14.62	14.62	13.58	-	
136	0.29	0.22	0.22	17.33	18.79	18.79	-	
137	0.19	0.17	0.16	19.42	19.83	20.04	-	
138	0.33	0.32	0.25	16.50	16.71	18.17	-	
139	0.26	0.22	0.22	17.96	18.79	18.79	-	
140	0.40	0.35	0.37	15.04	16.08	15.66	-	
141	0.18	0.22	0.08	19.62	18.79	21.71	-	
142	0.18	0.16	0.14	19.62	20.04	20.46	-	
143	0.11	0.10	0.07	21.08	21.29	21.92	-	
144	0.19	0.14	0.09	19.42	20.46	21.50	-	

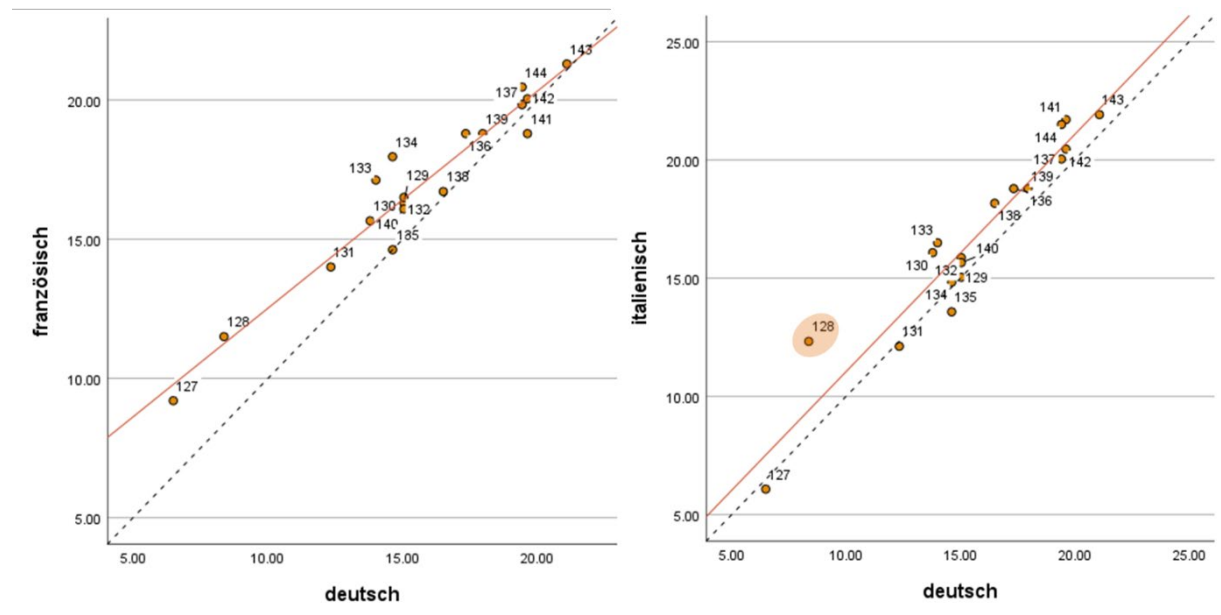


Abbildung 13: Delta-Plot für *Diagramme und Tabellen* (deutsch-französisch und deutsch-italienisch).

2.3 Effekte der Korrektur

Die durch den Ausgleich resultierenden Bonuspunkte können Tabelle 10 entnommen werden. Für die Bestimmung des Punktwertes wird auf ganze Punkte gerundet. Keine der Sprachgruppen erhält einen Sprachbonus, die Bonuspunkte kommen aber bei der Ermittlung der mittleren Rangplätze zum Tragen.

Tabelle 10: Mittelwerte und Standardabweichungen der Korrekturwerte für die Personen der französisch- und italienischsprachigen Gruppen.

		n	Minimum	Maximum	m	s
französisch	Textverständnis	347	0.00	0.12	0.08	0.06
	Fakten lernen	347	0.00	0.20	0.16	0.08
	GESAMT (Gerundet)	337	0.00	0.00	-	-
Italienisch	Fakten lernen	148	0.00	0.15	0.10	0.07
	Diagramme und Tabellen	148	0.00	0.19	0.09	0.09
	GESAMT (Gerundet)	139	0.00	0.00	-	-

Die Mittelwertsdifferenzen für jede Aufgabengruppe nach dem Sprachausgleich sind in Tabelle 11 dargestellt. Die Ergebnisse multipler Mittelwertvergleiche können der ersten Spalte entnommen werden. Differenzen zwischen den Teilnehmenden sind dort in der Form $d^{**} > (f, i)$ angegeben. Das Beispiel würde bedeuten, dass die deutschsprachigen Kandidaten und Kandidatinnen signifikant höhere Werte erzielt haben als die beiden anderen Sprachgruppen, die sich untereinander nicht signifikant unterscheiden.

Der Sprachbonus sowohl der französischen wie der italienischen Sprachgruppe war 2023 insgesamt geringer als in den Vorjahren (Abbildung 14).

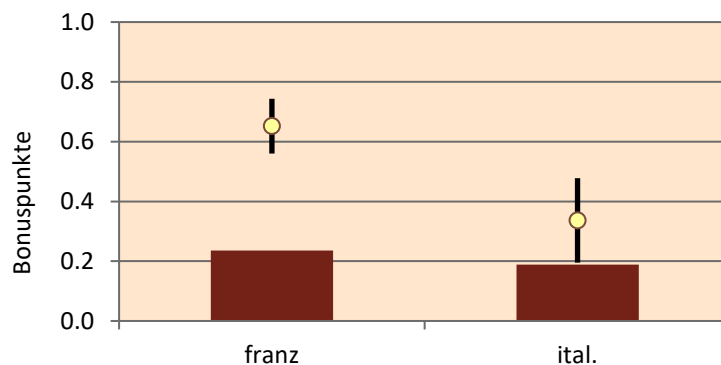


Abbildung 14: Sprachausgleich 2023 (ungerundet) mit Mittelwert und Standardabweichung der Testdurchführungen 2013-2022

Tabelle 11: Test-Prozentrang und Punktwerte inklusive Sprachausgleich.

		Mittelwert	Stand.- abw.	Minimum	Maximum
Test-Prozentrang d** > f	deutsch	52.04	28.11	10	100
	französisch	43.93	27.02	10	99
	italienisch	48.93	28.17	10	98
Quantitative und form.Probleme	deutsch	8.56	2.96	0	18
	französisch	8.27	3.11	1	18
	italienisch	8.36	2.94	1	16
Schlauchfiguren d**, i* > f	deutsch	11.93	3.49	1	18
	französisch	10.66	3.65	2	18
	italienisch	11.57	3.71	3	18
Textverständnis l* > d, f	deutsch	6.96	2.82	0	17
	französisch	6.76	2.51	1	15
	italienisch	7.55	2.79	1	16
Med.-naturwiss. Grundverständnis	deutsch	8.40	2.73	1	17
	französisch	8.61	2.64	1	17
	italienisch	8.34	2.71	1	15
Figuren lernen d* > f	deutsch	10.66	3.43	0	18
	französisch	10.15	3.45	2	18
	italienisch	10.22	3.33	2	17
Fakten lernen d > f**, i*	deutsch	9.93	3.69	0	18
	französisch	9.08	3.13	3	18
	italienisch	9.21	3.36	2	18
Muster zuordnen d**, i* > f	deutsch	9.01	3.18	0	18
	französisch	8.26	3.02	1	18
	italienisch	9.01	3.03	1	17
Diagramme und Tabellen d** > (f, i)	deutsch	6.73	2.81	0	18
	französisch	5.60	2.43	0	14
	italienisch	5.89	2.60	1	13
Konz. und sorgf. Arbeiten	deutsch	9.41	4.08	0	18
	französisch	9.32	3.83	0	18
	italienisch	9.58	3.97	0	18

n: deutsch 3305 / französisch 337 / italienisch 139

**/*: signifikanter Einfluss des Faktors „Sprache“ bei varianzanalytischer Prüfung (1%- bzw. 5%-Niveau); darunter signifikante Unterschiede bei multipltem Mittelwertvergleich.

Die Differenzen zwischen der deutschsprachigen Referenzgruppe und den anderssprachigen Teilnehmenden werden nach folgender Formel berechnet:

$$\text{Delta } z = \frac{m_{\text{Deutsch}} - m_{\text{Sprachgruppe}}}{s_{\text{Deutsch}}}$$

m ist der Mittelwert und s die Standardabweichung der entsprechenden Sprachgruppe. Die resultierenden Werte „Delta z“ drücken die Abweichung der jeweiligen Sprachgruppe von der deutschen Sprachgruppe in Standardabweichungen aus. 0.5 hiesse, dass die betroffene Sprachgruppe um eine halbe Standardabweichung geringere Punktzahlen erzielt als die deutschsprachige Referenzgruppe.

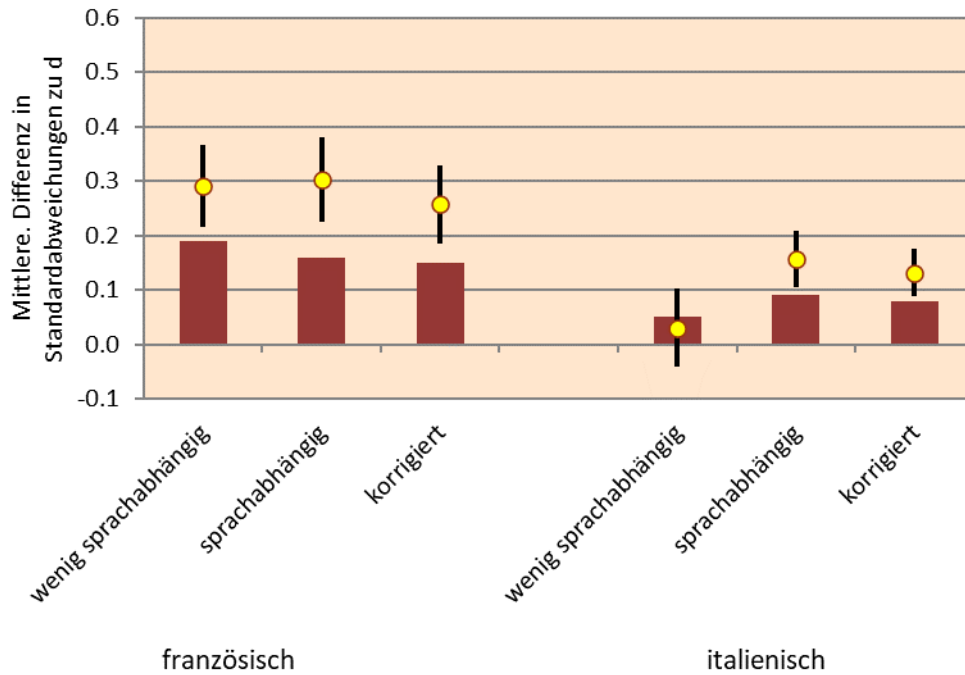


Abbildung 15: Mittlere Abweichung (Delta z) für wenig sprachabhängige und sprachabhängige Aufgabengruppen (korrigiert und unkorrigiert), 2023 und Vergleich zu 2013-2022 (Mittelwert und Streuung).

Die französische Sprachgruppe weist wiederum bei den sprachunabhängigen Aufgabengruppen eine höhere Differenz aus als bei den stärker sprachabhängigen. Nach der Korrektur wird letztere weiter vermindert. Die Abweichungen der italienischsprachigen Personen von der deutschen Sprachgruppe liegen bei den wenig sprachabhängigen Aufgabengruppen im (tiefen) Bereich der Vorjahre, bei den sprachabhängigen fallen sie geringer aus.

2.4 Vergleichbarkeit der Testlokale

2023 erfolgte die Testabnahme in 32 Testlokalen/Sektoren verteilt auf acht Testorte. Eine Besonderheit stellen wiederum die sehr kleinen Testlokale (Lokale 4 und 24) dar, die für Personen mit spezifischen Einschränkungen bereitgestellt wurden. Diese beiden Lokale werden aufgrund der zu geringen Fallzahlen von den weiteren Analysen ausgeschlossen.

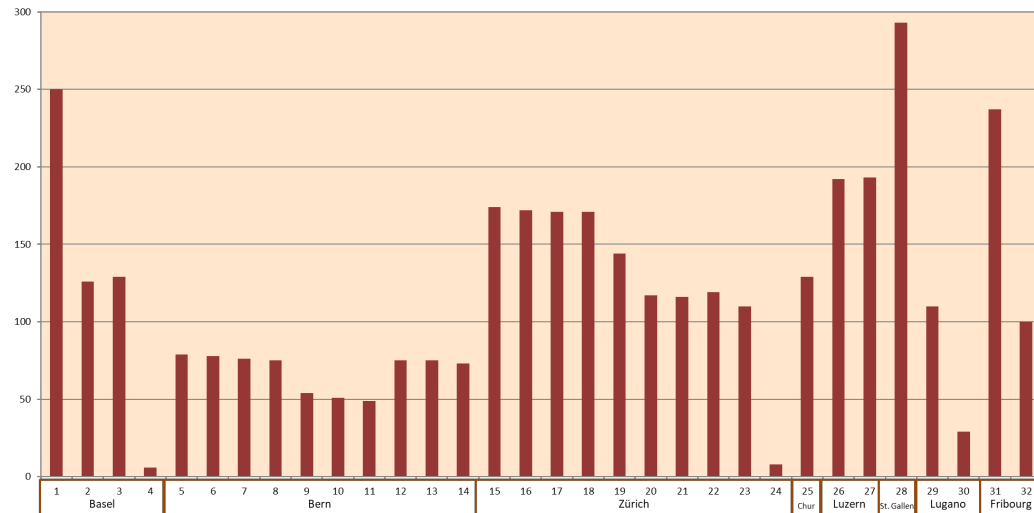


Abbildung 16: Testlokale mit Teilnehmerzahlen.

Tabelle 12: Varianzanalytische Prüfung der Homogenität für Test- und Punktwerte zwischen den deutschsprachigen Testlokalen (ohne Lokale 4 und 24).

		Quadratsumme	df	Mittl. Quadrat. Abweichung	F	Sig.
Testprozentrang	Zwischen	18801.149	25	752.046	0.952	0.531
	Innerhalb	2579016.477	3265	789.898		
	Total	2597817.627	3290			
Muster zuordnen	Zwischen	312.233	25	12.489	1.239	0.191
	Innerhalb	32908.731	3265	10.079		
	Total	33220.963	3290			
Med.-naturwiss. Grundverständnis	Zwischen	105.670	25	4.227	0.564	0.960
	Innerhalb	24467.974	3265	7.494		
	Total	24573.643	3290			
Schlauchfiguren	Zwischen	326.971	25	13.079	1.075	0.362
	Innerhalb	39709.509	3265	12.162		
	Total	40036.480	3290			
Quant. und formale Probleme	Zwischen	546.517	25	21.861	2.532	0.000
	Innerhalb	28185.190	3265	8.633		
	Total	28731.707	3290			
Textverständnis	Zwischen	248.304	25	9.932	1.257	0.177
	Innerhalb	25806.401	3265	7.904		
	Total	26054.706	3290			
Figuren lernen	Zwischen	114.736	25	4.589	0.388	0.997
	Innerhalb	38592.879	3265	11.820		
	Total	38707.615	3290			
Fakten lernen	Zwischen	300.848	25	12.034	0.885	0.629
	Innerhalb	44419.941	3265	13.605		
	Total	44720.788	3290			
Diagramme und Tabellen	Zwischen	170.448	25	6.818	0.865	0.657
	Innerhalb	25741.324	3265	7.884		
	Total	25911.772	3290			
Konzentr. und sorgf. Arbeiten	Zwischen	623.404	25	24.936	1.500	0.053
	Innerhalb	54271.267	3265	16.622		
	Total	54894.672	3290			

Zwischen den (deutschsprachigen) Testlokalen treten kaum signifikante Differenzen auf. Die varianzanalytische Prüfung weist lediglich für eine Aufgabengruppe signifikante Differenzen aus. Ein multipler Mittelwertsvergleich nach Tukey identifiziert lediglich zwei Testlokale mit geringeren Werten in einer Aufgabengruppe. Insgesamt sind die Unterschiede marginal.

Die Durchführungsbedingungen an den einzelnen Testorten sind standardisiert – aus den Testdurchführungsprotokollen ergeben sich keine Hinweise auf mögliche Ursachen vereinzelt auftretender Unterschiede. Es gibt keinen Anlass, Korrekturen aufgrund der Testlokalzuordnung vorzunehmen.

Tabelle 13: Darstellung der detaillierten Ergebnisse nach Testlokalen (ohne Lokale 4 und 24)

Lokal	Testprozentrang	Muster zuordnen	Med.-naturwiss. Grundverständnis	Schlauchfiguren	Quant. und formale Probleme	Textverständnis	Figuren lernen	Fakten lernen	Diagramme und Tabellen	Konzentr. u. sorgf. Arbeiten	Lokalgrösse
1	51.8	9.2	8.4	11.8	8.2	6.6	10.7	10.0	7.0	9.2	250
2	55.0	9.3	8.7	12.3	8.8	7.4	11.0	10.3	7.0	8.9	126
3	48.3	9.0	8.3	11.4	7.7	7.0	10.6	10.1	6.5	8.6	129
5	48.7	8.5	8.5	11.6	7.9	6.9	10.6	9.7	6.5	9.4	79
6	53.2	8.9	8.8	11.6	8.4	7.4	10.5	9.7	6.9	9.9	78
7	49.0	8.9	8.0	11.5	8.1	6.2	10.9	9.6	6.6	9.5	76
8	52.0	8.6	8.7	12.0	8.3	6.9	10.7	9.2	6.4	9.8	75
9	49.3	8.7	7.9	11.5	8.1	6.9	10.6	10.0	6.6	9.6	54
10	49.5	8.0	8.5	11.9	8.7	6.9	10.8	9.7	6.5	9.2	51
11	48.6	8.4	8.4	11.9	7.7	7.3	10.6	9.8	6.6	8.2	49
12	53.7	9.2	8.2	12.4	8.6	6.4	11.0	10.3	6.8	9.8	75
13	49.1	8.7	8.1	11.8	8.5	6.4	10.5	9.9	6.3	9.5	75
14	50.0	9.0	8.1	11.7	8.0	6.9	10.6	10.2	6.4	9.8	73
15	53.1	8.8	8.4	12.1	8.7	6.8	10.8	10.0	6.8	9.8	174
16	56.2	9.8	8.2	12.6	8.7	7.1	10.8	10.5	7.0	9.7	172
17	55.8	9.2	8.6	12.3	9.2	7.3	10.7	9.9	6.7	10.2	171
18	54.1	8.9	8.5	12.1	8.9	7.0	10.9	10.1	6.7	9.8	171
19	54.1	9.0	8.3	12.4	9.1	7.4	10.7	9.5	6.7	9.8	144
20	54.9	9.4	8.6	11.8	8.7	7.2	11.0	10.1	6.9	9.6	117
21	54.9	8.9	8.6	11.9	8.8	7.3	10.3	9.8	7.3	10.2	116
22	51.4	9.0	8.4	12.2	8.9	7.0	10.4	10.0	6.6	9.1	119
23	49.6	8.8	8.3	11.8	8.6	7.0	10.4	9.4	6.9	9.1	110
25	51.6	9.3	8.4	11.9	9.2	6.9	10.4	10.2	6.9	8.9	129
26	49.9	9.0	8.3	11.8	8.2	6.8	10.6	10.1	6.6	9.0	192
27	50.2	9.1	8.5	11.6	8.2	6.9	10.7	10.0	6.4	8.9	193
28	50.7	8.8	8.4	11.7	8.8	6.8	10.5	9.5	6.7	9.4	293
29	49.6	9.1	8.5	11.5	8.4	7.6	10.1	9.4	5.7	9.5	110
30	46.3	8.8	7.8	11.9	8.3	7.4	10.8	8.1	6.1	9.7	29
31	44.6	8.2	8.5	10.9	8.3	6.7	10.2	8.9	5.6	9.6	237
32	42.4	8.3	8.9	10.1	8.2	6.7	10.0	9.0	5.6	8.8	100
Total	51.2	8.9	8.4	11.8	8.5	7.0	10.6	9.8	6.6	9.4	3767

Markiert: Mittelwerte für Aufgabengruppen im multiplen Mittelwertsvergleich (Tukey), die signifikant tiefer als in mindestens einem anderen deutschsprachigen Testlokal liegen. Italienisch- und französischsprachige Durchführungen (Lokale 29 bis 32) sind farblich abgesetzt.

2.5 Vergleich für die Geschlechter

Die Forderung nach Gleichbehandlung bedeutet, dass bei gleicher Eignung die gleichen Chancen auf eine Zulassung bestehen. Identische Zulassungsquoten für die Geschlechter dürfen folglich bereits aufgrund der unterschiedlichen Repräsentativität der beiden Gruppen, die sich in einem deutlichen Ungleichgewicht der Teilnehmendenzahlen äussert, nicht erwartet werden.

Nach 2006 hatte sich der Männeranteil in Humanmedizin vorübergehend bei rund 40% stabilisiert, seit 2014 ist er allerdings weiter gesunken. 2023 blieb er gegenüber dem Vorjahr stabil, 31% der Bewerbenden mit gültigem Testergebnis waren männlich, 69 % weiblich (über alle Disziplinen).

Tabelle 14: Statistiken für den Test-Prozentrang nach Geschlechtern für Sprache und Disziplin.

		Testsprache			Wunschdisziplin			Gesamt
		deutsch	französisch	italienisch	HM	VM	ZM	
männlich	m	55.6	49.0	52.7	56.2	47.0	41.2	55.0
	s	28.5	29.4	27.2	28.4	25.9	27.5	28.5
	n	1040	73	61	1057	56	61	1174
weiblich	m	50.4	42.5	45.9	51.6	42.8	37.7	49.5
	s	27.8	26.2	28.7	27.7	26.2	26.3	27.8
	n	2265	264	78	2085	357	165	2607

m: Mittelwert, s: Standardabweichung, n: Personenzahl

Für die zahlenmässig bedeutendste deutschsprachige Teilnehmergruppe wird 2023 ein Unterschied von 5.2 Test-Prozentrangpunkten registriert. Die Differenz ist höher als 2022 (4.3), aber geringer als 2021 (6.5).

Tabelle 15: Kombinierte Statistiken für den Test-Prozentrang nach Geschlechtern, Sprache und Disziplin.

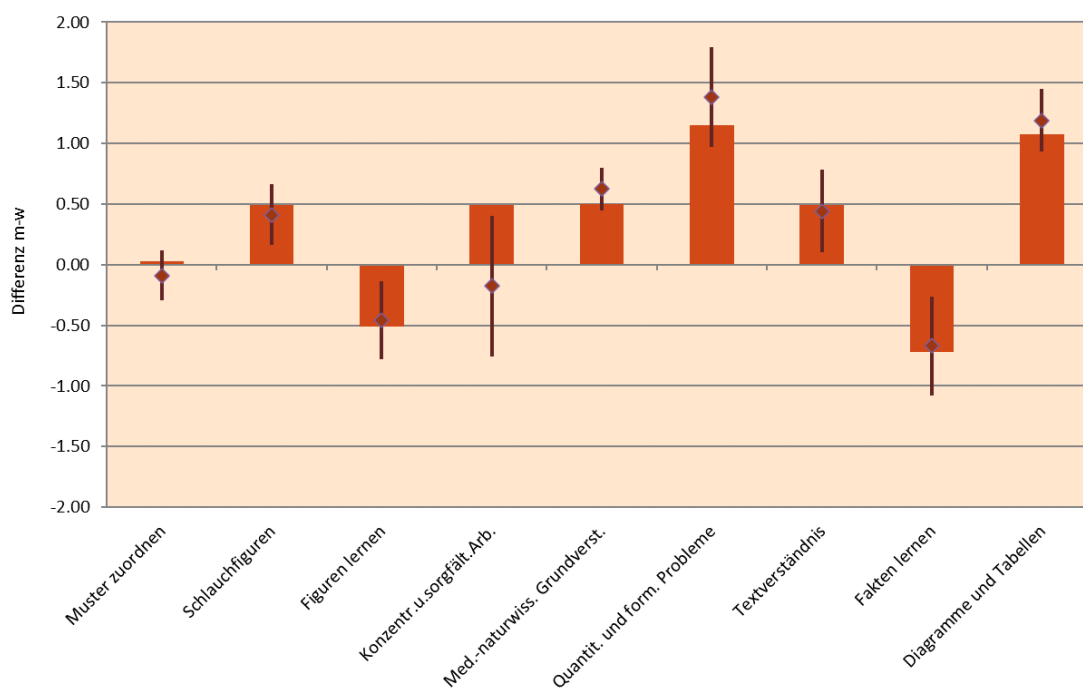
		Humanmedizin			Veterinärmedizin			Zahnmedizin		
		d	f	i	d	f	i	d	f	i
männlich	m	56.6	50.4	55.5	49.4	44.4	28.0	42.2		26.5
	s	28.4	30.7	26.7	26.4	25.1	25.5	28.0		15.7
	n	947	55	55	36	18	2	57		4
weiblich	m	52.1	48.4	45.1	46.1	34.2	54.6	37.9	17.0	26.0
	s	27.7	26.9	28.6	26.6	22.7	30.7	26.4	0.0	0.0
	n	1860	157	68	242	106	9	163	1	1

Betrachtet man die Unterschiede disziplin- und sprachspezifisch, zeigt sich für die deutschsprachige Gruppe mit Wunschdisziplin Humanmedizin eine signifikante Differenz von 4.5 Prozentrangpunkten. In der italienischen Sprachgruppe beträgt die Differenz dort 10.4 Prozentrangpunkte. Die anderen Unterschiede sind nicht signifikant.

Tabelle 16: Prüfung zur Varianzhomogenität und der Mittelwertunterschiede (t-Test) für Test-Prozentrang zwischen Geschlechtern (nur für ausreichende Zellengrößen).

Sprache	Disziplin	Levene-Test Gleichheit der Varianzen			t-Test Gleichheit des Mittelwerts		
			F	Sig.	t	df	Sig.
Deutsch	HM	homogen	0.89	0.35	4.09	2805	0.000
	VM	homogen	0.02	0.89	0.68	276	0.248
	ZM	homogen	1.11	0.29	1.04	218	0.150
Französisch	HM	homogen	2.94	0.09	0.47	210	0.319
Italienisch	HM	homogen	0.32	0.58	2.07	121	0.020

Abbildung 17 verdeutlicht die Vergleichbarkeit der Geschlechterdifferenzen mit den Vorjahren. Die diesjährigen Punktwert-Unterschiede auf Ebene der Aufgabengruppen fallen gegenüber früheren Durchführungen vergleichbar aus, sind einzig bei *Konzentriertes und sorgfältiges Arbeiten* etwas höher als üblich.



(positive Werte bedeuten höhere Ergebnisse für die männliche, negative Werte für die weibliche Stichprobe)

Abbildung 17: Geschlechtsspezifische Differenzen nach Aufgabengruppen (deutschsprachig, Humanmedizin) mit Vergleich (Mittelwert und Standardabweichung) für die Vorjahre 2013-2022.

Unter den gemäss Modell zugelassenen Personen (ohne Nachrücker) die 2023 am Test teilgenommen haben, beträgt der Frauenanteil in Humanmedizin 61.8%, in Veterinärmedizin 85.0%, in Zahnmedizin 72.3%, in Chiropraktik 82.4%.

2.6 Vergleiche nach Wunschhochschulen

Die Tabelle 17 führt die Test-Prozentränge getrennt nach Disziplin und gewünschter Hochschule (erste Wahl) aller deutschsprachigen Kandidaten und Kandidatinnen auf. Für die Bewerber und Bewerberinnen der Humanmedizin sind signifikante Differenzen (nach Tukey) zwischen Freiburg und den anderen Universitäten feststellbar, in der Veterinärmedizin haben Teilnehmende mit der Wunschuniversität Zürich bessere Ergebnisse erzielt.

Tabelle 17: Test-Prozentränge und Ergebnisse der Varianzanalyse pro Disziplin nach Wunschhochschulen (d).

	Wunschhochschule	n	Mittelwert	Standardabw.	Quadratsumme	Df	MQ	F	Sig.
HM	Basel	543	53.2	27.0	13481.8	6	2247.0	2.8	.010
	Bern	817	52.1	27.6					
	ETH	192	57.2	28.5					
	Freiburg	37	68.1	25.9					
	Zürich	1050	53.9	28.7					
	Zürich (LU)	73	55.2	30.4					
	Zürich (SG)	93	52.2	26.4					
VM	Bern	121	42.3	25.1	1291.2	1	1291.2	1.9	0.174
	Zürich	157	49.8	27.3					
ZM	Basel	54	36.5	28.2	3706.2	2	1853.1	2.9	0.060
	Bern	69	35.9	24.9					
	Zürich	97	42.7	27.1					

Unter den zugelassenen Personen treten zwischen den Wunschuniversitäten in der Humanmedizin keine signifikanten Unterschiede auf (Tabelle 18), die Differenz zwischen Bern und Zürich bleibt in der Veterinärmedizin signifikant.

Tabelle 18: Test-Prozentränge und Ergebnisse der Varianzanalyse pro Disziplin nach Wunschhochschulen, nur zugelassene*, deutschsprachige Personen.

	Wunschhochschule	n	Mittelwert	Standardabw.	Quadratsumme	Df	MQ	F	Sig.
HM	Basel	202	82.5	9.9	1003.0	6	167.2	1.8	.106
	Bern	290	82.6	9.5					
	ETH	82	85.4	9.9					
	Freiburg	23	85.3	9.5					
	Zürich	415	83.9	9.9					
	Zürich (LU)	34	83.3	9.7					
	Zürich (SG)	32	81.4	9.4					
VM	Bern	57	65.5	14.1	2154.9	1	2154.9	11.1	0.001
	Zürich	80	73.5	13.8					
ZM	Basel	22	64.9	22.1	416.4	2	208.2	0.5	0.602
	Bern	29	60.1	19.3					
	Zürich	59	60.0	19.9					

*Zulassung nach erstem, provisorischem Zulassungsmodell (ohne Nachrücker)

3 Ergebnisse zur Testgüte

3.1 Zuverlässigkeit

Von besonderem Interesse bleiben unter dem Blickwinkel der Neustrukturierung des Tests 2022 die Ergebnisse zur Testgüte. Aufgrund der vorangegangenen Modellierung wären hinsichtlich der Kennwerte keine auffälligen Auswirkungen der Kürzung im Vergleich zur Originalversion des EMS zu erwarten gewesen.

Die zulassungsrelevante Reliabilität (Zuverlässigkeit) über den Gesamtest kann anhand zweier Koeffizienten berechnet werden. Die Reliabilitätsschätzung nach der Testhalbierungsmethode (Teilung nach gerad- und ungeradzahigen Aufgaben) ist eine der gebräuchlichsten Zuverlässigkeitsschätzungen. Wegen der Stichprobengrößen werden nur die Ergebnisse des deutschsprachigen Tests der Schweiz verwendet.

Entscheidend sind die Kennwerte des Punktwertes, welcher nach der Standardisierung als Test-Prozentrang für die Zulassung verwendet wird.

Beide Werte sind, neben der Qualität der eigentlichen Aufgaben, auch abhängig von der Aufgabenzahl (im Fall der Konsistenz des Testprofils der Anzahl Aufgabengruppen). Nach zwei Jahren coronabedingter Kürzungen umfasste der Test 2022 wieder mehr Aufgaben, maximal waren statt 118 nun 162 Punkte möglich (die maximale Punktzahl des Originaltests lag bei 178 Punkten).

Der Zuverlässigkeitswert des Punktwertes erreicht 2023 mit .88 einen Wert im üblichen («vor-Corona») Bereich, das Testprofil mit 0.81 weiterhin eine gute Konsistenz. Beide Werte liegen damit im üblichen Bereich des Originaltests. Dieser Bereich scheint deshalb optimal, weil vergleichbare Werte auch in den Jahren mit einer erfolgreichen Evaluation des Zusammenhanges von Studienerfolg und Eignungstest gefunden worden sind.

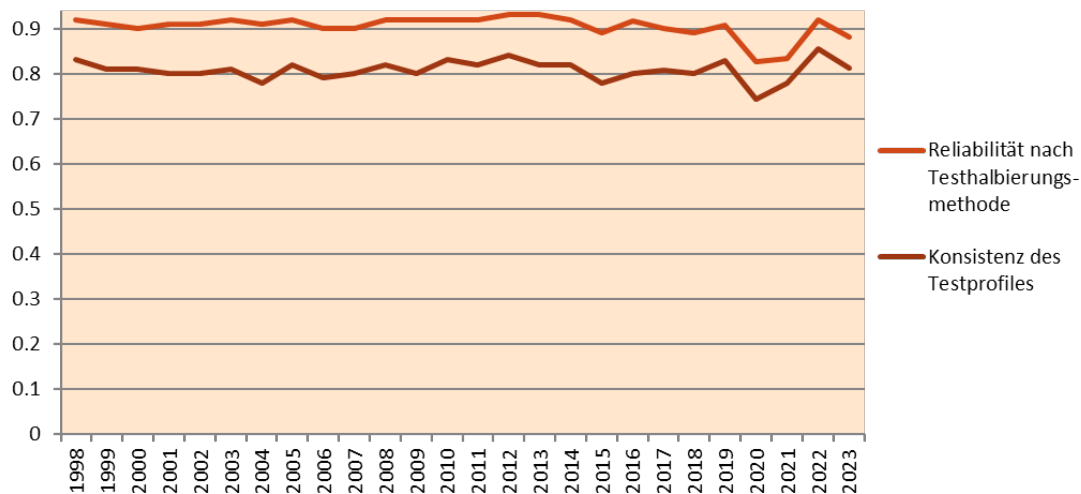


Abbildung 18: Reliabilität des Punktwertes (Split Half) und Konsistenz des Testprofils 1998-2023.

Die Zuverlässigkeitswerte der einzelnen Aufgabengruppen sind für die Testverwendung nicht bedeutsam, da keine Entscheidungen auf der Basis einzelner Aufgabengruppen getroffen und diese nicht einzeln interpretiert werden. Hier müssten lediglich Eingriffe erfolgen, wenn der Test-Prozentrang selbst eine zu geringe Zuverlässigkeit aufweisen würde. Die einzelnen Aufgaben werden im Zug der Testauswertung auf die Eindeutigkeit der Fragestellung hin geprüft und bezüglich Schwierigkeit und Trennschärfe analysiert. 2023 mussten keine Aufgaben aus Qualitätsgründen aus der Wertung entfernt werden. Die verwendeten

Aufgabengruppen weisen die geforderten Varianzen und Schwierigkeiten auf, um zwischen den Kandidierenden optimal zu differenzieren.

3.2 Binnenstruktur

Aufgrund der Korrelationen zwischen den Aufgabengruppen kann mittels Strukturanalyse geprüft werden, ob sich die einzelnen Aufgabengruppen bestimmten Dimensionen zuordnen lassen. Die so gewonnene Struktur bietet Vergleichsmöglichkeiten mit theoretischen Vorstellungen zum Fähigkeitsbereich. Die Faktorenstruktur kann zu den bekannten Faktormodellen der Intelligenz in Beziehung gesetzt werden.

Tabelle 19: Korrelationen zwischen Punktwerten der Aufgabengruppen und Gesamt-Punktwert.

	Med.-naturwis. Grundv.	Schlauchfiguren	Quant. u. formale Probl.	Textverständnis	Figuren lernen	Fakten lernen	Diagr. und Tabellen	Konzent. u. sorgf. Arbeiten	Punkt-wert
Muster zuordnen	.240**	.503**	.242**	.207**	.375**	.481**	.268**	.177**	.613**
Med.-naturw. Grundverst.		.341**	.466**	.502**	.281**	.311**	.493**	.286**	.652**
Schlauchfiguren			.342**	.294**	.434**	.462**	.389**	.287**	.715**
Quant. u. form. Probleme				.426**	.261**	.253**	.533**	.330**	.648**
Textverständnis					.240**	.271**	.461**	.242**	.603**
Figuren lernen						.475**	.296**	.177**	.626**
Fakten lernen							.311**	.157**	.662**
Diagramme und Tabellen								.360**	.691**
Konzent. u. sorgf. Arbeiten									.552**

Die Varianzanteile der unrotierten Lösung für die sechs statt neun Aufgabengruppen entsprechen den bekannten Grössenordnungen (Tabelle 20).

Tabelle 20: Faktorenanalyse: Verlauf der unrotierten Lösung (2019-2023).

Faktor	Unrotierte Lösung														
	Eigenwert					% Varianz					% Kumuliert				
	2019	2020	2021	2022	2023	2019	2020	2021	2022	2023	2019	2020	2021	2022	2023
1	3.84	2.73	2.95	4.17	3.74	42.66	45.52	49.17	46.37	41.51	42.66	45.52	49.17	46.37	41.51
2	1.40	0.94	0.97	1.14	1.28	15.51	15.63	16.09	12.68	14.23	58.17	61.15	65.26	59.06	55.74
3	0.71	0.71	0.60	0.72	0.82	7.93	11.86	9.94	8.01	9.07	66.10	73.01	75.20	67.06	64.81
4	0.63	0.67	0.57	0.61	0.63	7.05	11.17	9.55	6.74	7.05	73.16	84.17	84.75	73.81	71.86
5	0.56	0.52	0.48	0.57	0.60	6.27	8.60	7.97	6.34	6.62	79.43	92.77	92.72	80.14	78.48
6	0.52	0.43	0.44	0.53	0.52	5.75	7.23	7.28	5.89	5.75	85.17	100	100	86.03	84.22
7	0.50			0.44	0.50	5.56			4.91	5.53	90.73			90.94	89.75
8	0.43			0.42	0.47	4.82			4.63	5.25	95.55			95.57	95.00
9	0.40			0.40	0.45	4.45			4.43	5.00	100			100	100

Tabelle 21: Varianzanteile der 2-Faktorenlösung 2019-2023 (rotiert).

Varimax-rotierte Lösungen														
Eigenwert					% Varianz					% Kumuliert				
2019	2020	2021	2022	2023	2019	2020	2021	2022	2023	2019	2020	2021	2022	2023
2.7	2.3	2.4	2.7	2.7	29.8	37.7	40.7	30.0	29.5	29.8	37.7	40.7	30.0	29.5
2.6	1.4	1.5	2.6	2.4	28.3	23.5	24.6	29.1	26.2	58.2	61.2	65.3	59.1	55.7

Ein allgemeiner Fähigkeitsfaktor klärt 46% der Varianz auf – es gibt einen grossen „gemeinsamen Nenner“ in allen Aufgabengruppen. Mit der Zwei-Faktorenlösung können 56% der Varianz aufgeklärt werden. Die Aufgabe *Konzentriertes und sorgfältiges Arbeiten* ist je nach konkreter Aufgabenstellung eher dem einen oder dem anderen Faktor zuzuordnen – misst aber in jedem Fall (wie intendiert) eine durch die anderen Aufgabengruppen weniger berücksichtigte Fähigkeit. 2023 ist diese eher dem Faktor „schlussfolgerndes Denken“ zuzuordnen.

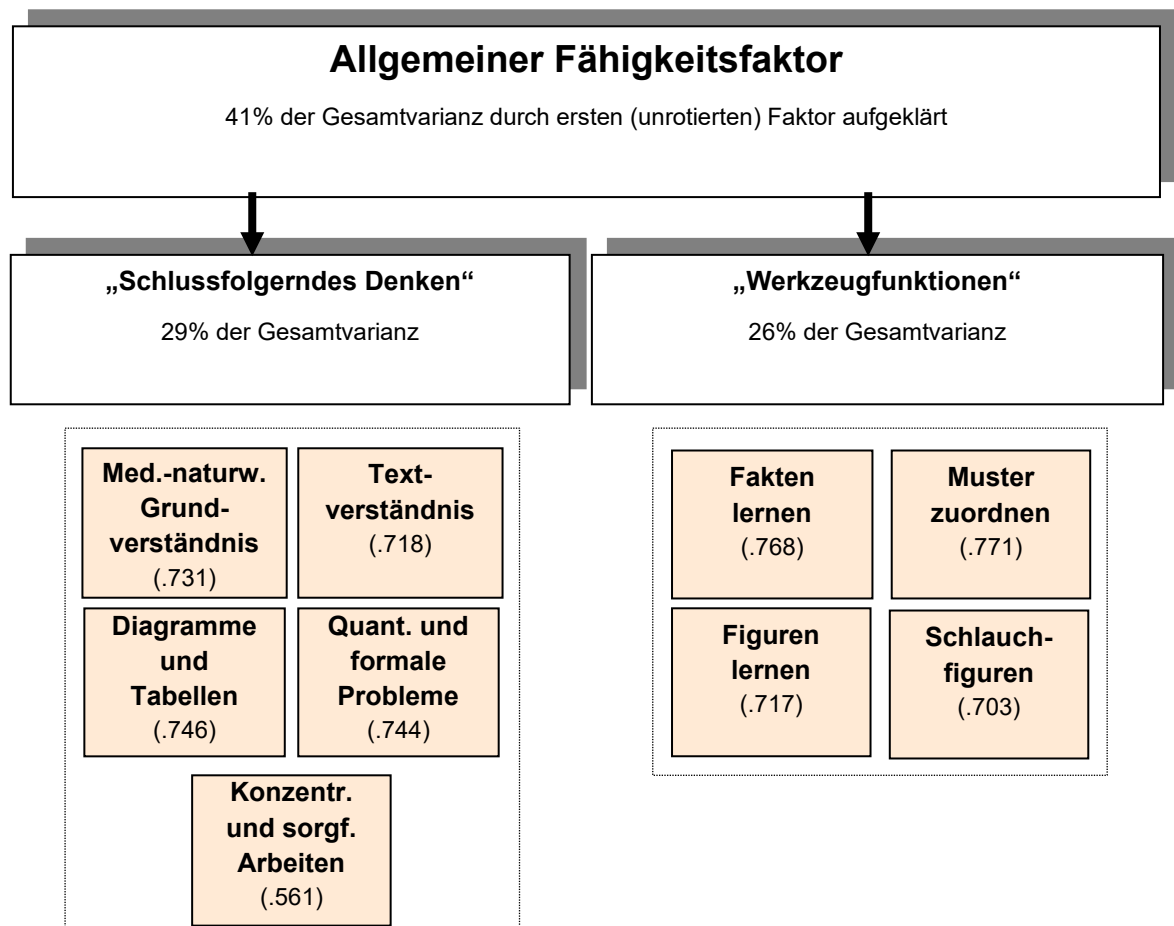


Abbildung 19: Struktur der Aufgabengruppen des EMS, hierarchische Faktorenstruktur 2023 mit Faktorladungen der Aufgabengruppen auf rotierten Faktoren.

Die aktuellen Werte entsprechen durch den wiederhergestellten Faktor „Werkzeugfunktionen“ den Ergebnissen vor den Corona-Jahren bis 2019. Insgesamt hatte die neue Teststruktur keine nachteiligen Auswirkungen auf die Testgüte.

4 Literatur

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In Holland, P.W., Wainer, H. (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Beller, M., Gafni, N. (1995). Translated Versions of Israel's interuniversity Psychometric Entrance Test (PET). In T. Oakland & R.K. Hambleton (Eds.), *International Perspectives of Academic Assessment*, S.207-218. Boston: Kluwer.

Beller, M. (1996). Translating, equating and validating Scholastic Aptitude Tests: The Israeli Case. In Hänsgen, K.-D., Hofer, R. & Ruefli, D. (Hrsg). *Eignungsdiagnostik und Medizinstudium*, (S. 14-29), Bericht 2. Freiburg/Schweiz: Zentrum für Testentwicklung und Diagnostik.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (Second Edition)*. Hillsdale, NJ: Erlbaum.

Cook, L.L. (1998). Can Scores Obtained on Test Given in Different Languages to Examinees of Different Cultures be Equally Valid? ICAP: San Francisco.

Deidesheimer Kreis (1997). Hochschulzulassung und Studieneignungstests: studienfeld-bezogene Verfahren zur Feststellung der Eignung für Numerus clausus und andere Studiengänge. Göttingen, Zürich: Vandenhoeck und Ruprecht.

Ebach, J., Trost, G. (1997). *Admission to Medical Schools in Europe*. Lengerich: Pabst.

Hänsgen, K.-D, (2014): Wollen wir nicht lieber diplomierte Ärzte? Sind fachliche und soziale Kompetenzen Gegensätze? *Schweizerische Ärztezeitung* 2014; 95: 6 S. 194 f.

Hänsgen, K.-D. (2013): Ohne Numerus clausus sinkt die Qualität des Studiums. Der externe Standpunkt, *NZZ* am Sonntag, 28.4. 2013, S. 17.

Hänsgen, K.-D (2007): Numerus clausus in der Medizin – werden die Richtigen ausgewählt für Studium und Beruf? *Schweizerische Ärztezeitung | Bulletin des médecins suisses | Bollettino dei medici svizzeri* | 2007; 88: 46.

Hänsgen, K.-D. (2007): Wie bilden wir die richtigen Studierenden zu Ärzten aus? *Neue Zürcher Zeitung*, 18. Oktober 2007.

Hänsgen, K.-D., Spicher, B. (2002). Numerus clausus: Finden wir mit dem «Eignungstest für das Medizinstudium» die Geeigneten? *Schweizerische Ärztezeitung / Bulletin des médecins suisses / Bollettino dei medici svizzeri* – 2002; 83(31): 1653 – 1660. <http://www.saez.ch/pdf/2002/2002-31/2002-31-842.PDF>

Hänsgen, K.-D., Spicher, B. (2002). Numerus clausus : Le « test d'aptitudes pour les études de médecine » (AMS) permet-il de trouver les personnes les plus aptes? *Schweizerische Ärztezeitung / Bulletin des médecins suisses / Bollettino dei medici svizzeri* – 2002; 83 (47) 2562 – 2569. <http://www.saez.ch/pdf/2002/2002-47/2002-47-1144.PDF>

Hänsgen, K.-D., Spicher, B. (2000). Zwei Jahre Numerus clausus und Eignungstest für das Medizinstudium in der Schweiz (EMS). Teil 1: Erfahrungen mit dem EMS als Zulassungskriterium. *Schweizerische Ärztezeitung* Heft 12, S. 666 – 672.

Hänsgen, K.-D., Spicher, B. (2000). Zwei Jahre Numerus clausus und Eignungstest für das Medizinstudium in der Schweiz (EMS). Teil 2: EMS und Chancengleichheit. *Schweizerische Ärztezeitung* Heft 13, S. 723-730.

Hängsen, K.-D., Hofer, R. & Ruefli, D. (1995a). Un test d'aptitudes aux études de médecine est-il faisable en Suisse? Bulletin des médecins suisses, 7, S. 267 - 274.

Hängsen, K.-D., Hofer, R. & Ruefli, D. (1995b). Der Eignungstest für das Medizinstudium in der Schweiz. Schweizerische Ärztezeitung, 37, S. 1476 - 1496.

Longford, Holland & Thayer, (1993). Stability of the MH D-DIF Statistics Across Populations. In P. W. Holland & H. Wainer (Ed.) Differential Item Functioning, S. 171 – 196. Lawrence Erlbaum Associates: New Jersey.

Oswald, U. (1999). Der Eignungstest 1998 für das Medizinstudium. Schweizerische Ärztezeitung 80, S. 1313 – 1317.

Trost, G. (Hrsg.) (1994). Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (18. Arbeitsbericht). Bonn: ITB.

Trost, G. (Hrsg.) (1995). Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (19. Arbeitsbericht). Bonn: ITB.

Trost, G. (Hrsg.) (1996). Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (20. Arbeitsbericht). Bonn: ITB.

Trost, G. (Hrsg.) (1997). Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (21. Arbeitsbericht). Bonn: ITB.

Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M. & Nauels, H.-U. (1998). Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn: ITB.

4.1 Frühere Berichte des ZTD

Über EMS-Seite: <http://www.unifr.ch/ztd/de/ems/berichte.html>