UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

ZTD ZENTRUM FÜR TESTENTWICKLUNG UND DIAGNOSTIK AM DEPARTEMENT FÜR PSYCHOLOGIE
CENTRE POUR LE DÉVELOPPEMENT DE TESTS ET LE DIAGNOSTIC AU DEPARTEMENT DE PSYCHOLOGIE
UNI Rte d'Englisberg 9 1763 GRANGES-PACCOT ☎ +41 263007986 🖷 +41(26)3009763 www.unifr.ch/ztd

Zusatzbericht EMS 2014: Nichtwertung der 14 Aufgaben

Inhalt

Zusammenfassung

*Beim diesjährigen Eignungstest für das Medizinstudium (EMS) hat es einen **möglichen Vorteil** für eine unbestimmte Personenzahl durch die vorherige Bekanntheit von 14 Aufgaben gegeben. Deren Ausschluss aus der Wertung war notwendig für die Wiederherstellung der Gleichbehandlung. Die vom Gesetz vorgeschriebene Feststellung der Eignung ist auf der Basis des gewerteten Teils gewährleistet, die Nichtwertung entspricht den Qualitätsstandards für Psychodiagnostik: Psychologische Tests erfordern bestmögliche Bedingungsgleichheit für alle Personen.*
Obwohl die Daten keine Hinweise auf eine bedeutsame, qualitativ oder quantitativ identifizierbare Personengruppe mit Nutzen aus diesem Vorwissen liefern, wurden weitere mögliche Vor-/Nachteile für bestimmte Teilnehmergruppen theoretisch und empirisch geprüft:

- *Ein bedeutsamer Vorteil als **Zeitgewinn** bei der Bearbeitung der gewerteten Aufgaben für die Personen mit Vorkenntnis ist aufgrund der empirischen Daten und wissenschaftlicher Kenntnisse zur Diagnostik des „Schlussfolgernden Denkens" nicht nachweisbar.*
- *Die Veränderung der **Gewichte** der einzelnen Aufgabengruppen am Gesamtergebnis liegt im Schwankungsbereich der bisher eingesetzten Varianten EMS. Auch der für die Studienerfolgsprognose gewichtige Faktor „Schlussfolgerndes Denken" ist noch ausreichend repräsentiert.*
- *Die **Äquivalenz** (Ersetzbarkeit) der Testergebnisse aufgrund der ursprünglichen und der gekürzten Punktsummen ist für die drei betroffenen Aufgabengruppen sowie insbesondere für den Gesamttest ausreichend hoch. Die Reihung der Personen als Basis für die Zulassung bleibt zwischen beiden Versionen sehr gut vergleichbar.*

*Bei **„Fakten lernen"** besteht nur eine sehr hohe Ähnlichkeit der Lernliste, die dazu gestellten Fragen sind andere. Die Ergebnisse zeigen, dass Personen, die am wahrscheinlichsten an einem Vorbereitungskurs teilgenommen haben, im Mittel sogar schlechtere Leistungen erzielen als andere. Dies ist psychologisch aufgrund des Konzeptes der Interferenz erklärbar – die Nichtwertung ist bei einer differenzierten Güterabwägung nicht notwendig. Die gewertete Version des EMS 2014 erfüllt alle Anforderungen für eine gültige Testabnahme und Zulassung aufgrund der gemessenen Eignung.*

Was ist geschehen – welche Massnahmen wurden geprüft und ergriffen?

Ende Juli uns übersandte Unterlagen eines Trainingsanbieters enthalten 14 mit dem EMS 2014 identische Aufgaben aus 3 Aufgabengruppen: 1 Text der Aufgabengruppe „Textverständnis" mit 6 Fragen, sowie je 4 Aufgaben „Quantitative und Formale Probleme" und „Diagramme und Tabellen". Bei der Aufgabengruppe „Fakten lernen" gab es Ähnlichkeiten zwischen Trainingsmaterial und Original: Die Lernliste war sehr ähnlich, es wurden aber in der Abrufphase andere Fragen dazu gestellt.
Dieses Trainingsmaterial entspricht einer in Deutschland verwendeten Originalversion des TMS aus dem Jahr 1996, die so nie in der Schweiz eingesetzt wurde. Es handelt sich um ein Original oder eine entsprechende Kopie (nicht abfotografiert, rekonstruiert o.ä.). Es gibt keine Hinweise auf weitere im Umlauf befindliche Originalaufgaben. Sonstige am Markt verfügbare rekonstruierte oder nachempfundene Aufgaben der Trainingsanbieter sind damit qualitativ nicht vergleichbar – aufgrund der Marktbeobachtung in der Schweiz und auch in Deutschland gab es bisher keine Gründe, andere Originalaufgaben zurückzuziehen.

Die Schweiz adaptiert seit 1998 den deutschen TMS (drei Sprachen, Anpassung an Schweizer Gegebenheiten). Im Rahmen der kontinuierlichen Erneuerung werden einzelne Aufgaben ersetzt – so sind die Aufgaben in die Version 2014 gelangt und deshalb sind es auch nur wenige.
Die vorher bekannten identischen Aufgaben, die von einer unbekannten Personenzahl unter irregulären Bedingungen bearbeitet worden sind, wurden aus Sicherheitsgründen nicht gewertet. Die Bedingungen sind dadurch bei den gewerteten Aufgaben **hinsichtlich des Vorwissens** für alle Personen wieder gleich. Die Abbildung 1 stellt die weiteren notwendigen Prüf- und Entscheidungsschritte dar.

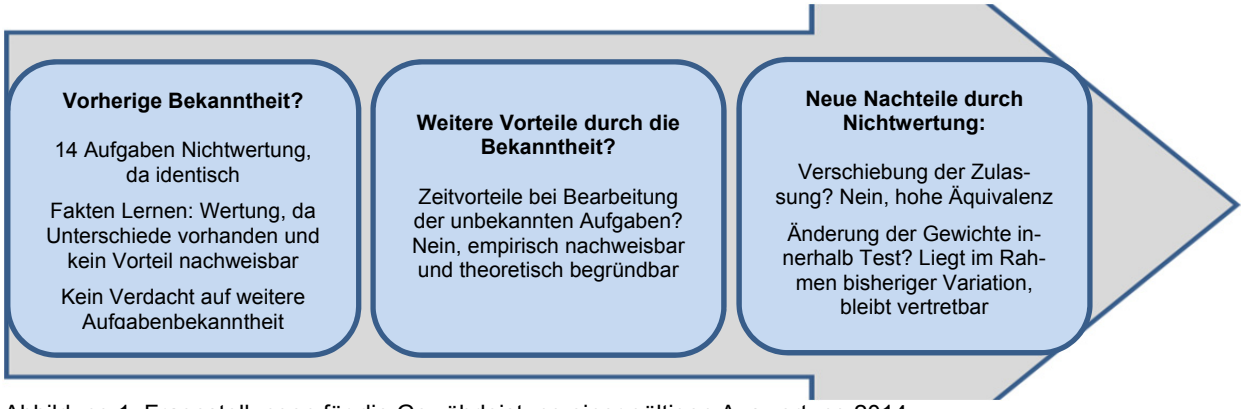


Abbildung 1: Fragestellungen für die Gewährleistung einer gültigen Auswertung 2014

Sind die bekannten Aufgaben auffällig hinsichtlich Schwierigkeit oder Trennschärfe?

Die vorher bekannten Aufgaben reihen sich bezogen auf Schwierigkeit (Lösungswahrscheinlichkeit) und Trennschärfe (Korrelation Lösungswahrscheinlichkeit mit Gesamtleistung der Aufgabengruppe) an den aufgrund der in Deutschland 1996 ermittelten Kennwerten in die jeweilige Aufgabengruppe ein, sind z.B. nicht im Mittel leichter oder mit niedrigerer Trennschärfe verbunden, weil z.B. die Beantwortung anders funktioniert hätte (Tabelle 1). Sie sind über diese vergleichsweise lange Zeit gleich schwer geblieben, was für hohe zeitliche Stabilität der gemessenen Merkmale spricht.

Tabelle 1: Schwierigkeiten (Lösungswahrscheinlichkeiten p) und Trennschärfen (Korrelation Aufgabe mit Gesamtwert r) sowie Standardabweichung (s) für betroffene Aufgabengruppen – ungewertete Aufgaben markiert

Quantitative und Formale Probleme Zeit: 50 Minuten				Diagramme und Tabellen Zeit: 50 Minuten				Textverständnis Zeit: 45 Minuten			
Nr.	p	s	r	Nr.	p	s	r	Nr.	p	s	r
1	.88	.325	.306	**159**	.83	.379	.207	**41**	.74	.436	.257
2	.80	.397	.294	**160**	.83	.377	.158	**42**	.73	.441	.252
3	.74	.440	.287	**161**	.78	.414	.303	**43**	.57	.495	.352
4	.78	.412	.297	**162**	.71	.456	.374	**44**	.59	.492	.299
5	.71	.455	.350	**163**	.69	.463	.375	**45**	.63	.482	.398
6	.64	.480	.316	**164**	.61	.489	.174	**46**	.23	.420	.261
7	.77	.418	.337	**165**	.67	.470	.253	**47**	.77	.424	.220
8	.69	.462	.330	**166**	.52	.500	.298	**48**	.60	.490	.290
9	.54	.498	.272	**167**	.65	.478	.352	**49**	.67	.472	.295
10	.69	.464	.355	**168**	.50	.500	.314	**50**	.55	.497	.202
11	.66	.475	.318	**169**	.59	.491	.308	**51**	.39	.487	.325
12	.65	.476	.304	**170**	.47	.499	.304	**52**	.51	.500	.236
13	.57	.496	.348	**171**	.53	.499	.312	**53**	.73	.442	.278
14	.53	.499	.347	**172**	.53	.499	.330	**54**	.49	.500	.281
15	.46	.498	.290	**173**	.42	.493	.285	**55**	.58	.494	.302
16	.42	.493	.359	**174**	.30	.459	.270	**56**	.41	.492	.307
17	.48	.500	.308	**175**	.36	.479	.264	**57**	.33	.469	.329
18	.35	.477	.273	**176**	.26	.440	.180	**58**	.27	.445	.314
19	.31	.463	.240	**177**	.29	.453	.293	Schwierigkeiten hier innerhalb der 3 Texte mit je 6 Fragen abgestuft			
20	.29	.454	.163	**178**	.22	.416	.188				

Gibt es Auffälligkeiten in der Verteilung der Punktwerte?

Wenn nennenswerte Teilgruppen einzelne Aufgaben unter irregulären Bedingungen (z.B. vorherige Bekanntheit) beantwortet haben, müsste sich dies als Auffälligkeit der Antwortverteilung zeigen. Häufungen bei hohen Punktzahlen (bis zu Mehrgipfligkeiten) sprächen für qualitative Unterschiede. Ist der Unterschied eher quantitativ (z.B. nur geringere relative Verbesserungen bei diesen Personen), müsste sich dies als Schiefe zeigen, weil diese Teilgruppe bessere Werte als erwartet erreicht.
Das Maximum von 14 Punkten in den 14 vorher bekannten Aufgaben (sehr gute Personen oder Personen, die optimal von Vorwissen profitierten) erreichen nur 33 Personen. Dies sind 1% der Gesamtgruppe – was die Erwartung auch unter normalen Umständen wäre. Fasst man 13 und 14 Punkte zusammen, sind es 109 Personen, was ebenfalls in der Erwartung liegt. Gäbe es eine nennenswerte Gruppe von Personen, die (fast) alle Aufgaben aufgrund der vorherigen Bekanntheit gelöst hat, müssten hier mehr Personen vertreten sein. Abbildung 2 zeigt, dass sowohl in den gewerteten (unbekannten), als auch in den ungewerteten (möglicherweise bekannten) Aufgaben die Punkteverteilung nahezu ideal einer symmetrischen Glockenverteilung folgt und auch keine Schiefe als Zeichen quantitativer Unterschiede vorhanden ist (Schiefe .00 für gewertete, .08 für ungewertete Aufgaben, Kurtosis bzw. Wölbung: -.50 und -.54). Entweder haben sehr wenige Personen dieses Material tatsächlich gesehen, oder diese haben sich die richtigen Lösungen der vorher bekannten Aufgaben nicht merken können.

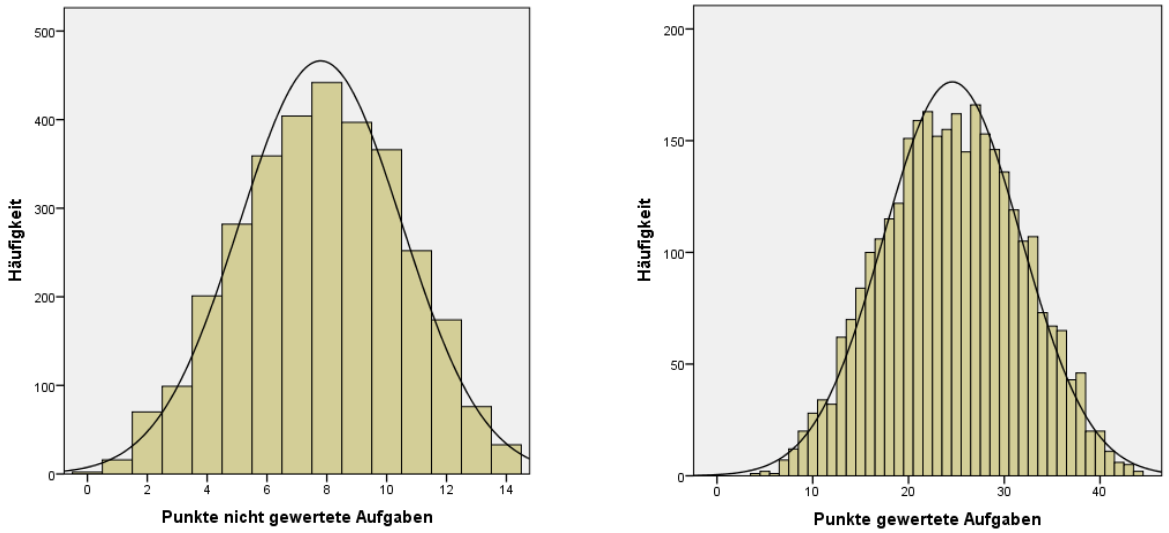


Abbildung 2: Vergleich der Punktwerteverteilung für nicht gewertete und gewertete Aufgaben der drei betroffenen Aufgabengruppe

Hat die Bekanntheit Einfluss auf die Bearbeitung unbekannter Aufgaben?

Es wurden Vorteile von Personen mit Vorkenntnissen der 14 Aufgaben auch bei den gewerteten Aufgaben der betroffenen drei Aufgabengruppen vermutet, etwa durch gewonnene Zeit. Die betroffenen Aufgabengruppen gehören allerdings zu jenen mit vergleichsweise reichlicher Bearbeitungszeit, weil dort kein zu grosser Zeitdruck gewünscht ist (2 x 50 und 1 x 45 Minuten).
Trotz fehlendem generellen Nachweis von Vorteilen bei der Verteilungsanalyse soll dem nachgegangen werden: Abbildung 3 liefert zunächst eine „visuelle“ Übersicht über die Daten. Dort wurden die Personen nach dem EMS-Gesamt-Punktwert in fünf ca. 20% umfassende Leistungs-Gruppen eingeteilt. Analog wurden fünf Klassen nach der Punktzahl nur für die ungewerteten Aufgaben gebildet, in der besten Klasse sind Personen mit 13 und 14 Punkten zusammengefasst. Gewertete und ungewertete Aufgaben sind durch den Bezug auf den gleichen Messbereich hoch korreliert, das Niveau in beiden sollte auch vom Gesamt-Punktniveau abhängen.
Die Boxplots zeigen für die Gruppen nach Punkten in den bekannten Aufgaben ein homogenes Bild: Sie steigen **zwischen** den gleichen Punktwerteklassen ungewerteter Aufgaben **innerhalb** jeder Gesamt-Punktwerteklasse sowie auch zwischen den Klassen der Gesamt-Punktwerte gleichförmig. Auch die „beste“ Klasse mit 13 und 14 Punkten in den bekannten Aufgaben setzt sich von den übrigen nicht ab, bei einem substantiellen Punktezuwachs durch Zeitgewinn müssten sich die Klassen mit den höchsten Punktzahlen in den ungewerteten Aufgaben „unabhängiger“ von ihrem Gesamtpunktwert darstellen und bessere Ergebnisse erreichen.

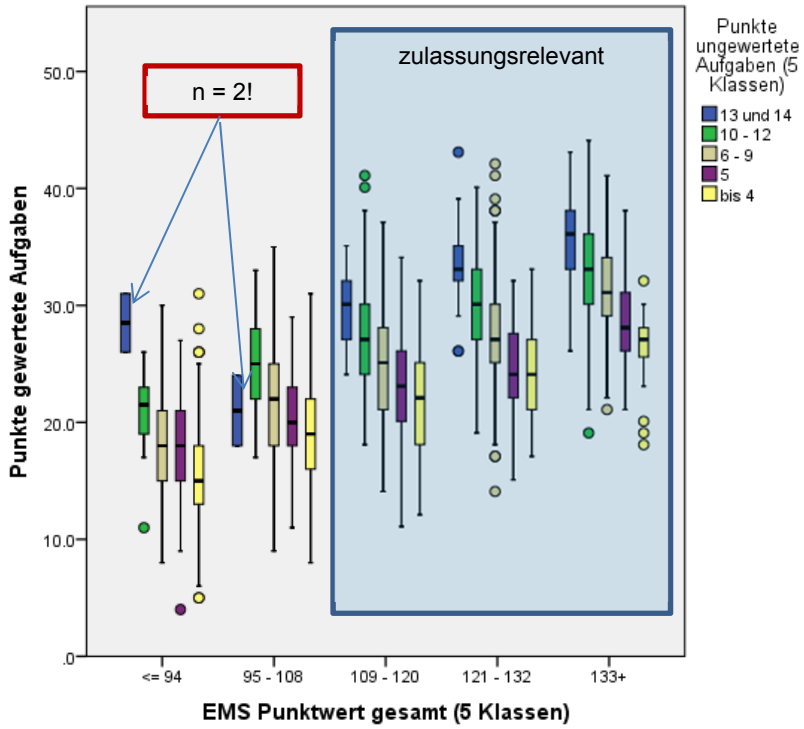


Abbildung 3: Boxplots (Mediane und Quartilabstände) der Punktsumme der gewerteten Aufgaben der drei Aufgabengruppen Diagramme und Tabellen, Textverständnis, Quantitative und formale Probleme für je 5 Punktwerteklassen für Gesamt und 5 Klassen Punkte nur in den 14 ungewerteten Aufgaben. 2 Klassen sind nur mit 2 Personen besetzt!
Punkte: Ausreisser

Eine genaue Identifikation jener Personen, die von der Bekanntheit der Aufgaben profitiert haben können, ist nicht realistisch, da Unterschiede ebenso messfehlerbedingt sein können. Für die Anwendung statistischer Prüfmethoden sollen stattdessen Personengruppen mit der **höchsten Wahrscheinlichkeit** eines Profitierens von der Aufgabenbekanntheit herausgefiltert und verglichen werden:

- Relative Unterschiede: Ein Weg führte über eine Regressionsanalyse. Über die Punktwerte in allen gewerteten Aufgaben des Faktors „Schlussfolgerndes Denken“ kann abgeschätzt werden, welcher Punktwert in den 14 eliminierten Aufgaben aufgrund des individuellen Leistungsniveaus jedes Kandidaten erwartet wird. 50 Personen, deren Punktsummen diesen Erwartungswert am deutlichsten übertrafen, wurden in die „**Trainingsgruppe**“ eingeteilt – dort ist die höchste Wahrscheinlichkeit gegeben, dass diese Personen von der Bekanntheit der Aufgaben profitiert haben. Nicht alle diese Personen müssen am Trainingskurs teilgenommen haben, der Unterschied kann auch zufällig bzw. messfehlerbedingt sein. Wenn es aber überhaupt einen systematischen Effekt eines Profitierens gibt, der grösser als der Messfehler ist, müssen sich die „trainierten“ Personen in dieser Gruppe gehäuft finden. Andernfalls hätte die Bekanntheit zu überhaupt keinen relevanten Vorteilen geführt. Diese Trainingsgruppe unterscheidet sich nicht von den übrigen Personen bezüglich der mittleren Gesamtleistung (113.2 Vergleichsgruppe, 113.5 Trainingsgruppe als mittlerer Punktwert) – die Differenz zwischen gewerteten und ungewerteten Aufgaben kann auf allen Leistungsstufen auftreten.
- Absolute Unterschiede: Eine zweite **„Extremgruppe 90/50“** wird aus 49 Personen gebildet, die in den 14 ungewerteten Aufgaben mindestens einen Prozentrang von 90 erreicht haben (zu den besten 10% in diesen Aufgaben gehören, mindestens 11 Aufgaben gelöst haben), sowie im Faktor „Schlussfolgerndes Denken“ über alle gewerteten Aufgaben maximal den Prozentrang 50 erreicht haben (zu den „schlechteren“ 50% in den gewerteten Aufgaben dieses Faktors gehören). Auch hier ist die Diskrepanz zwischen gewerteten und ungewerteten Aufgaben am höchsten – es gilt die gleiche Einschränkung wie oben, dass die Unterschiede auch messfehlerbedingt sein können. Diese Extremgruppe hat bedingt durch die Auswahl durchschnittlich schlechtere Gesamtleistungen als die übrigen Personen – die Leistung in den ungewerteten Aufgaben ist „absolut“ besser.

Vergleicht man die mittleren Lösungswahrscheinlichkeiten für gewertete und ungewertete Aufgaben innerhalb der drei betroffenen Aufgabengruppen, wird deutlich (Abbildung 4): Die durchschnittlichen Lösungswahrscheinlichkeiten beider Personengruppen mit dem wahrscheinlichsten Vorwissen bewegen sich in allen drei Untertests für die bekannten Aufgaben zwischen 0.8 und 0.9 (was durch die Art der Gruppenbildung trivial ist). Bedeutsam ist, dass bei den unbekannten Aufgaben aller drei betroffenen Aufgabengruppen diese Personengruppen jeweils gleiche oder etwas geringere mittlere Punktwerte aufweisen als die Vergleichsgruppe. Das Ausmass dieser Unterschiede entspricht im Niveau genau den Differenzen in beiden nicht betroffenen Aufgabengruppen des gleichen Faktors „Schlussfolgerndes Denken“.

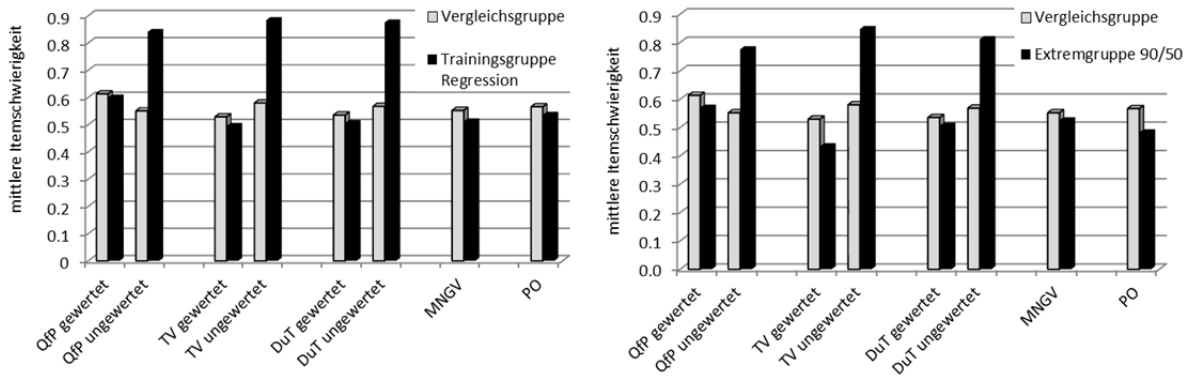


Abbildung 4: Vergleich mittlerer Schwierigkeiten gewertete und nicht gewertete Aufgaben für Gruppen aufgrund Regression und Prozenträngen; QfP: Quantitative und formale Probleme, TV: Textverständnis, DuT: Diagramme und Tabellen, MNGV: Medizinisch-Naturwissenschaftliches Grundverständnis. PO: Planen und Organisieren

Wenn die Bekanntheit einiger Aufgaben der betroffenen Aufgabengruppen einen Einfluss auf die Bearbeitung der übrigen Aufgaben derselben Aufgabengruppe hätte (z.B. durch Zeitgewinn), **müsste die Trainingsgruppe in den gewerteten Aufgaben der betroffenen Aufgabengruppen tendenziell besser abschneiden als die Vergleichsgruppe, und besser als in den beiden nicht betroffenen Aufgabengruppen dieses Faktors sein**. Die detaillierten Ergebnisse entsprechender Mittelwertsvergleiche (t-Test) für die betroffenen Aufgabengruppen sind in Tabelle 2 dargestellt und nur **für die nicht gewerteten Aufgaben wie erwartet signifikant.** (Die Signifikanz bei gewerteten Aufgaben ist wie beschrieben aufgrund der Art der Personenauswahl trivial).

Tabelle 2: Trainingsgruppe und Vergleichsgruppe (n=2780): Punktwerte in betroffenen Aufgabengruppen – signifikante Unterschiede nur bei den 14 ungewerteten Aufgaben, nicht bei den gewerteten. Signifikanzprüfung mittels t-Tests für unabhängige (vertikal) bzw. abhängige (horizontal) Stichproben

Gruppe	Mittlerer Punktwert		Signifikanz der Differenz beobachteter Wert vs. Erwartungswert	Mittlere Differenz zw. Erwartungswert und tatsächlichem Wert	Standardabweichung (beobachtet)
	beobachtet	erwartet aufgrund Regression			
Punktwert in den 14 ungewerteten Aufgaben					
Vergleichsgruppe	7.78	7.86	**.023***	-0.08	2.69
Trainingsgruppe	12.16	7.57	**.000****	4.59	1.63
Signifikanz Gruppenvergleich	**.000****	.282		**.000****	
Quantitative und formale Probleme					
Vergleichsgruppe	9.81	9.81	.948	-0.00	3.12
Trainingsgruppe	9.60	9.45	.668	0.15	3.15
Signifikanz Gruppenvergleich	.642	.237		.626	
Diagramme und Tabellen					
Vergleichsgruppe	8.56	8.56	.966	0.00	2.98
Trainingsgruppe	8.10	8.20	.741	-0.90	2.66
Signifikanz Gruppenvergleich	.277	.227		.750	
Textverständnis					
Vergleichsgruppe	6.34	6.34	.972	0.00	2.61
Trainingsgruppe	5.94	6.01	.775	-0.07	2.36
Signifikanz Gruppenvergleich	.280	.086		.792	

*: Signifikant auf dem 5%-Niveau; **: Signifikant auf dem 1%-Niveau

Für alle Vergleiche auf Ebene der Aufgabengruppen gilt: Sowohl der mittlere Punktwert wie auch die mittlere Differenz zum Erwartungswert unterscheiden sich **nicht signifikant**. Auch in der „Trainingsgruppe“ entsprechen die Punktwerte in den gewerteten Aufgaben den **durch das individuelle Leistungsniveau erwarteten Werten. Für Zeitvorteile durch die Bekanntheit der Aufgaben finden sich hier keine Belege.**

Dies wird auch durch theoretische Konzepte der Psychologie gestützt: Bei Aufgaben des Schlussfolgernden Denkens steht Zeit in keiner linearen Beziehung zur Leistung – begrenzend ist vielmehr die zeitlich stabilere Fähigkeit als dispositionelles Merkmal. Einige vergleichbare Tests („Niveau- oder

Powertests“) arbeiten ganz ohne Zeitbegrenzung und die Ergebnisse differenzieren dennoch nach der Fähigkeit. Übersteigt die Schwierigkeit einer Aufgabe die persönliche Fähigkeit, wird entweder geraten oder auch eine falsche Antwort gegeben. Die Personen kommen innerhalb der verfügbaren Zeit an die Grenze, bei welcher die Aufgabenschwierigkeit die eigene Fähigkeit übersteigt. Das Konzept der ansteigenden Schwierigkeit findet sich auch bei den drei betroffenen Aufgabengruppen.

Es wurde auch geprüft, ob **einzelne** weitere Aufgaben der Version 2014 in einem Trainingsmaterial des Anbieters enthalten und damit bekannt waren. Vergleiche einzelner Aufgabenschwierigkeiten der Trainingsgruppe mit den übrigen Personen zeigt, dass sich nur die 14 ungewerteten Aufgaben aus den betroffenen Aufgabengruppen wie erwartet signifikant unterscheiden (Abbildung 5). Würden einzelne weitere Aufgaben ebenfalls bekannt gewesen sein, müssten sie gleichfalls diesem Muster folgen. Diese Analyse wurde auch für alle anderen Aufgabengruppen des EMS 2014 vorgenommen – keine weitere Aufgabe weist derartige signifikante Differenzen auf. Dieser Befund bezieht sich auf diesen einen Trainingsanbieter. Hinweise auf weitere identische Aufgaben anderer Anbieter liegen nicht vor und sind aus verschiedenen Gründen auch weniger wahrscheinlich.

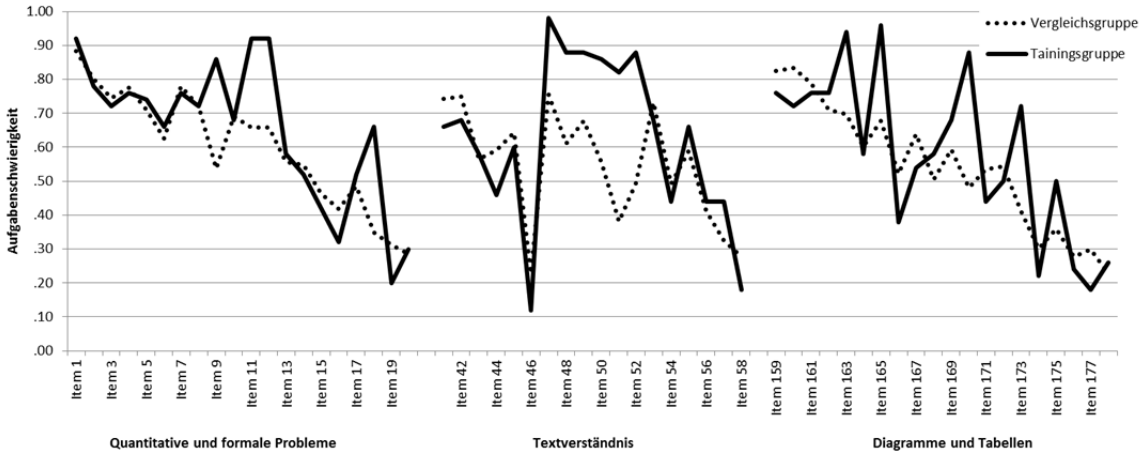


Abbildung 5: Schwierigkeiten der Aufgaben für die Trainingsgruppe im Vergleich zu den anderen Teilnehmenden

Zusammenfassend finden sich keine Hinweise dafür, dass eine Gruppe mit Vorkenntnissen der Aufgaben weitere Vorteile aus der Aufgabenbekanntheit gezogen hätte (etwa durch Zeitvorteile). Die Nichtwertung der Aufgaben hat die möglichen Vorteile korrigiert (selbst wenn auch deren Nachweis nicht wirklich evident erscheint).

Entstanden neue Nachteile dadurch, dass 14 Aufgaben nicht gewertet wurden?

Der Einwand, dass es Personen gibt, die in den 14 Aufgaben besser waren und sich durch die Nichtwertung **systematisch** verschlechtert hätten, gilt nur für jene, die von der Bekanntheit profitiert hätten. Die Leistungen in den gewerteten und ungewerteten Aufgaben nach Aufgabengruppen sind konstruktionsbedingt hoch korreliert. Die Rangreihe nach der Leistung muss sich gleichermassen zeigen, alle nicht systematischen Abweichungen sind aus Sicht der psychologischen Theorie Messfehler. Dies muss so sein, weil alle Einzelaufgaben das gleiche Merkmal messen – nicht wie bei Prüfungen inhaltlich unterschiedliche Facetten eines Themengebietes. Zwischen den Aufgaben gibt es keinen inhaltlichen Bezug, sondern nur formal gleiche Anforderungen (z.B. Textverständnis).

Tabelle 3: Äquivalenzschätzung der Punktwerte für gesamte und verkürzte Aufgabengruppen

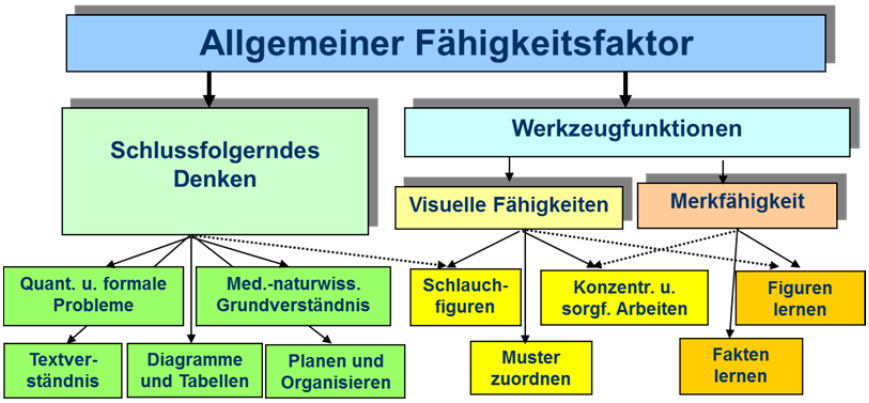
Gültige Version	Ursprüngliche Version	**Textverständnis** 18 Aufgaben	**Quantitative und formale Probleme** 20 Aufgaben	**Diagramme und Tabellen** 20 Aufgaben
Textverständnis (12 Aufgaben)	Korrelation Pearson	**.913****	.474**	.508**
	N	3173	3173	3173
Quantitative und formale Probleme (16 Aufgaben)	Korrelation Pearson	.499**	**.967****	.637**
	N	3173	3173	3173
Diagramme und Tabellen (16 Aufgaben)	Korrelation Pearson	.530**	.623**	**.964****
	N	3173	3173	3173

**: Signifikant auf dem 1%-Niveau

Die **Paralleltest-Reliabilität** prüft **die Äquivalenz von 2 Messungen**. Sind die Messungen mit allen und nur den gewerteten Aufgaben äquivalent, **kann eine Messung die andere ersetzen**, beide Messungen differenzieren die Personen dann auf gleiche Weise. Für die drei betroffenen Aufgabengruppen korrelieren die Summen der ursprünglichen mit den gewerteten Aufgaben mit 0.91 bis 0.97 (Tabelle 3). Reliabilitäten ab 0.85 werden als hoch eingeschätzt – da eine Teilmenge mit einem Ganzen verglichen wird, muss der Wert hier etwas höher sein. In allen drei Fällen liegt eine ausreichende Äquivalenz vor, die gewerteten Aufgaben messen jeweils das gleiche wie ursprünglich alle Aufgaben.
Entscheidender ist allerdings die Äquivalenz des zulassungsrelevanten Gesamttests: **Für den gesamten Test erreicht die Äquivalenz für den Punktwert sogar 0.996, für den Test-Prozentrang 0.995 und für den mittleren Rangplatz 0.996. Dies ist mehr als ausreichend für die Annahme einer vergleichbaren Rangreihe nach der Studieneignung und einer vergleichbaren Zulassung.**

Haben sich durch die Nichtwertung die Gewichte im EMS bedeutsam verändert?

Die drei betroffenen Aufgabengruppen wie auch die nicht betroffenen Aufgabengruppen „Medizinisch-Naturwissenschaftliches Grundverständnis“ und „Planen und Organisieren“ gehören zum wichtigen Faktor **„Schlussfolgerndes Denken“**. Dieser varianzstärkste Faktor ist normalerweise mit 98 Aufgaben vertreten, jetzt noch mit 84. Bis 2004 umfasste er nur 78 Aufgaben (wie noch heute im deutschen TMS). Die Gewichtung des Faktors ist also immer noch höher als in den ersten Formen des EMS.



Es wurde erwogen, die Resultate der betroffenen Aufgabengruppen proportional auf die vollen Punktzahlen hochzurechnen. Das wären aber „neue Spielregeln während des laufenden Spiels“ gewesen. Die Varianz wäre „künstlich“ erhöht worden, es hätten Zwischenwerte gefehlt und durch eine einzige Aufgabe hätte man plötzlich z.B. 1.2 Punkte erhalten oder verloren. Deshalb wurde davon abgesehen.

Gibt es Vorteile durch die Ähnlichkeit beim „Fakten lernen“?

Wirken sich die Unterschiede der Lernliste und die anderen dazu gestellten Fragen auf die Leistungen aus? Analog Abbildung 3 werden dazu in Tabelle 4 die Punktwerte gesamt und in den vorher bekannten Aufgaben hinsichtlich des Punktwerts im „Fakten lernen“ untersucht. In jeder Klasse der nicht gewerteten Aufgabenpunkte steigt der Mittelwert für „Fakten lernen“ mit steigendem Gesamtpunktwert erwartungsgemäss an, da die Leistungen auch korrelieren.
Bei den besten 60% (ab 109 Punkte) liegt das jeweilige Maximum des Mittelwerts (markiert) nie in der Klasse mit der höchsten Punktzahl in den nicht gewerteten Aufgaben (dies müsste der Fall sein, wenn diese Gruppe besonders profitiert hätte).

Tabelle 4: Mittelwerte (m) und Standardabweichungen (s) sowie Personenzahl (n) in der Aufgabengruppe „Fakten lernen“ für je 5 Punktwerteklassen und 5 Klassen Punkte in den 14 nicht gewerteten Aufgaben

Klassen Gesamt-punkte	Klassen Punkte in den 14 nicht gewerteten Aufgaben														
	13 und 14			10 bis 12			6 bis 9			5			4 und weniger		
	n	m	s	n	m	s	n	m	s	n	m	s	n	m	s
<= 94	2 (!)	*n.b.*	*n.b.*	20	**10.5**	2.7	179	**10.9**	3.6	109	**10.8**	3.5	337	**10.5**	3.8
95 - 108	2 (!)	*n.b.*	*n.b.*	69	**12.1**	3.2	278	**13.2**	3.3	111	**14.0**	3.2	163	**13.5**	3.8
109 - 120	9	**12.1**	4.0	166	**13.9**	3.4	331	**14.8**	3.5	61	**15.8**	2.8	93	**15.5**	3.3
121 - 132	26	**14.4**	4.3	224	**15.8**	3.4	290	**16.6**	2.8	52	**18.0**	2.2	49	**17.7**	2.4
133+	70	**17.6**	2.9	313	**17.7**	2.6	165	**18.1**	2.2	26	**18.6**	1.6	28	**18.5**	2.1

n.b: Nicht berechnet, da nur 2 Personen und nicht zulassungsrelevant

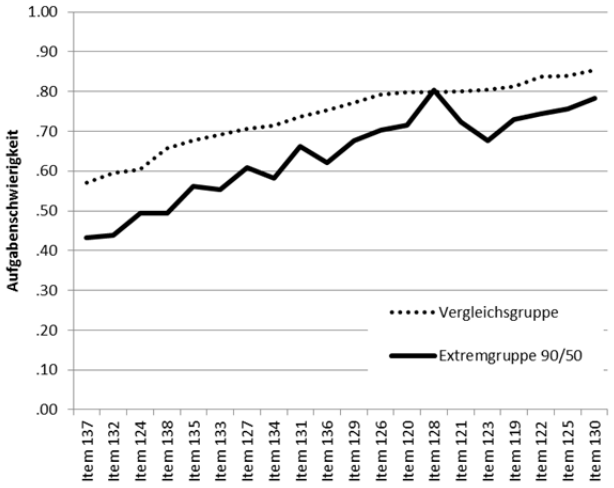


Abbildung 6: Vergleich der Schwierigkeiten der einzelnen Aufgaben bei „Fakten lernen“ für Extrem- und Vergleichsgruppe

Analog des auf Seite 3 beschriebenen Verfahrens der Bildung der „Extremgruppe 90/50“ wird für die Aufgabenanalyse von „Fakten lernen“ eine Extremgruppe aus 148 Personen gebildet, die mindestens 11 der 14 ungewerteten Aufgaben korrekt beantwortet haben und gleichzeitig maximal Prozentrang 50 im Faktor „**Werkzeugfunktionen**“ aufweisen (welchem „Fakten lernen“ zuzuordnen ist). Dies sind Personen, bei denen die Leistung bei „Fakten lernen“ am wahrscheinlichsten über der Erwartung liegt – mit gleichen Einschränkungen wie oben, dass dies auch messfehlerbedingt sein kann. Aus diesem Vergleich ergibt sich **kein Hinweis auf einen Vorteil dieser Extremgruppe** bei der Bearbeitung der Aufgabengruppe „Fakten lernen“. Wie erwartet sind diese Werte schlechter (Abbildung 6).

Weiterhin wird auch für die durch Regressionsanalyse ermittelte Trainingsgruppe (siehe Seite 4) deutlich, dass der Unterschied in der Leistung beim „Fakten lernen“ ebenfalls nicht signifikant ist (Tabelle 5) – für diese Gruppe unterscheiden sich die mittleren Gesamtleistungen bekanntlich ebenfalls nicht. Bei der Bildung dieser Gruppe ist „Fakten lernen“ nicht berücksichtigt worden, der fehlende signifikante Unterschied ist daher besonders aussagefähig.

Tabelle 5: Vergleichsgruppe und Trainingsgruppe aufgrund Regression für „Fakten lernen“. Die Signifikanzprüfung erfolgte mittels t-Tests für unabhängige (vertikal) bzw. abhängige (horizontal) Stichproben

Gruppe	**Mittlerer Punktwert**		Signifikanz der Differenz beobachteter Wert vs. Erwartungswert	Mittlere Differenz zw. Erwartungswert und tatsächlichem Wert	Standardabweichung (beobachtet)
	beobachtet	erwartet aufgrund Regression			
Fakten lernen (auch hier kein signifikanter Unterschied – für alle 20 Aufgaben)					
Vergleichsgruppe	14.69	14.70	.873	-0.01	4.10
Trainingsgruppe	15.50	14.93	.226	0.6	4.02
Signifikanz Gruppenvergleich	.167	.490		.229	

Die Aufgabengruppe gehörte von Beginn des EMS an zu den leichteren, wobei häufiger hohe bzw. volle Punktzahlen erreicht wurden und die Varianz der Punktwerte niedriger war. Insofern hätten sich Personen mit Vorkenntnissen auch überhaupt weniger von den übrigen absetzen können.
Da sich keine Vorteile für Personengruppen nach möglicher Vorkenntnis nachweisen lassen (sondern in einer Analyse sogar Nachteile) ist die Wertung und Beibehaltung dieser Aufgabengruppe im Rahmen einer Güterabwägung gerechtfertigt.

Eine mögliche psychologische Erklärung, warum Personen mit „Vorkenntnissen“ sogar etwas schlechter sind, sei kurz dargestellt: Das Umlernen einer vorhandenen Lernliste ist in der Regel schwererer als das Neulernen. Beim Abruf der gelernten Dinge kann es zu sogenannten proaktiven Interferenzen zwischen ursprünglicher und veränderter Liste kommen, indem diese nicht ausreichend unterschieden werden können. Die Vorgabe der ähnlichen Liste reduziert die Reproduktionsleistung der ursprünglichen Liste. Underwood (1957) hat dies in dem Satz zusammengefasst „Je mehr {Ergänzung: und auch je intensiver} Listen vorher gelernt wurden, desto geringer ist die Wahrscheinlichkeit, die neue Liste korrekt zu erinnern“. Danach haben Personen, die ähnliche Listen vorher gelernt haben, sogar Nachteile gegenüber Personen, die eine Liste neu lernen, was die Ergebnisse erklären würde.