*Univ.-Prof. Mag. Dr. Martin Arendasy, Mag. Dr. Markus Sommer*

*Universität Graz*

*Wolfgang M. Prodinger, Mag. Martina Heidegger*

*Medizinische Universität Innsbruck*

*Joachim Fritz Punter, Hon.-Prof. Dr. Markus Grimm, Univ.-Prof.in Dr.in Anita Rieder*

*Medizinische Universität Wien*

*Mag. Dr. phil.Sabine Vogl, Mag. Daniel Ithaler*

*Medizinische Universität Graz*

*Univ.-Prof. Mag. Dr. Stefan Koch*

*Johannes Kepler-Universität Linz*

# Aufnahmeverfahren MedAT: Fairness und inhaltliche Breite als zentrale Kriterien

# 1. Aufnahmeverfahren: Unde venistis

Im Jahre 2006 fanden in Österreich erstmals Aufnahmeverfahren statt, um über die Vergabe der bestehenden staatlich finanzierten Ausbildungsplätze in Human- und Zahnmedizin zu entscheiden. Der Ausganspunkt hierfür war ein Urteil des Europäischen Gerichtshofs vom 7.7.2005, das in Folge des Beitritts Österreichs zur EU im Jahre 1995 das Recht auf einen Studienplatz in Österreich auch für EU-Bürger durchsetzte. Dadurch kam es zu einer weiteren Überschreitung der damals ohnedies bereits sehr angespannten Ausbildungsressourcen vor allem aufgrund der hohen Nachfrage von StudienplatzwerberInnen aus Deutschland. Infolge der Novellierung des Universitätsgesetzes 2002 wurden die vom deutschen Numerus Clausus betroffenen Studiengänge dazu ermächtigt, Zugangsbeschränkungen zu erlassen und StudienplatzwerberInnen mit Hilfe von Aufnahmeverfahren auszuwählen.[1] Österreich war dabei eines der letzten Länder, die angesichts der hohen Nachfrage im Verhältnis zu den vorhandenen, staatlich finanzierten Ausbildungsressourcen den freien Zugang zum Medizinstudium beenden mussten, um die Qualität der Ausbildung auch weiterhin gewährleisten zu können.[2] Bereits 1997 gab es nur in 5 von 35 Europäischen Ländern einen freien Zugang zum Medizinstudium ohne Aufnahmeverfahren.[3]

1 Haag et al (2020), Grimm & Marschall (2016)
2 Ebach & Trost (1997)
3 Ebach & Trost (1997)

# 2. Wege zum Medizinstudium: Beispiele aus Europa

In der Schweiz bestand der Grund für die Einführung von Aufnahmeverfahren ebenfalls in einer zunehmend steigenden Anzahl an StudienplatzwerberInnen und den damit verbundenen Überschreitungen der verfügbaren Ausbildungsressourcen. Bei dem in der Schweiz verwendeten Aufnahmeverfahren handelt es sich um den Schweizer Eignungstest für das Medizinstudium (EMS), der in wesentlichen Teilen auf dem deutschen Test für Medizinische Studiengänge (TMS) basiert.[4]

In Deutschland werden Aufnahmeverfahren für das Medizinstudium in Kombination mit Schulnoten bereits seit längerer Zeit angewandt. Beispielsweise wurde der Test für medizinische Studiengänge (TMS) von 1986 bis 1996 bundesweit verpflichtend eingesetzt. Seit 2007 findet er wieder freiwillig als eines von mehreren möglichen Aufnahmekriterien statt. Ergänzend zu Schulnoten können aufnehmende Medizinische Universitäten auch noch auf den TMS, den Hamburger Naturwissenschaftstest (HAM-Nat), Multiple Mini-Interviews und künftig auch auf einen Situational-Judgment-Test zurückgreifen.[5] Die beiden letztgenannten Verfahren dienen der Erfassung sozialer und emotionaler Kompetenzen, während der HAM-Nat Kenntnisse in den naturwissenschaftlichen Fächern Biologie, Chemie, Physik und Mathematik erfasst. Darüber hinaus beinhaltet er auch noch eine Aufgabengruppe zum logischen Denken. Ein einheitliches bundesweit verpflichtendes Aufnahmeverfahren existiert jedoch nicht. Vielmehr können aufnehmende Medizinische Fakultäten und Hochschulen nach eigenem Ermessen aus der oben beschriebenen Toolbox als Ergänzung zu Schulnoten auswählen. Lediglich Wartelisten und Losverfahren wurden unlängst als verfassungswidrig erklärt.[6]

Der Fokus auf Schulnoten zeigt sich auch in anderen Ländern. Beispielsweise werden Schulnoten auch in Italien neben einem Aufnahmeverfahren, das in einigen Aspekten dem MedAT ähnlich ist, herangezogen, um über die Aufnahme zum Medizinstudium zu entscheiden. Ähnliches gilt auch in den Niederlanden, in denen neben Schulnoten auch naturwissenschaftliche Kenntnistests und Situational-Judgment-Tests bei der Entscheidung über die Aufnahme zum Medizinstudium berücksichtigt werden.

4 Hänsgen (2000a, b)
5 zusammenfassend: Hampe & Kadmon (2019); Schwibbe et al (2018); Trost (1989)
6 zusammenfassend: Hampe & Kadmon (2019); Schwibbe et al (2018); Trost (1989)

Der Fokus auf Schulnoten als einem von mehreren Aufnahmekriterien findet sich auch im Vereinigten Königreich. Hier wird jedoch zusätzlich auch ein nationales Aufnahmeverfahren, der University Clinical Aptitude Test (UCAT) angewandt.[7] Dieser umfasst neben Aufgabengruppen zu kognitiven Fähigkeiten auch eine Aufgabengruppe in Form eines Situational-Judgment-Tests, mit dessen Hilfe soziale und emotionale Kompetenzen erfasst werden, die dem entsprechenden Testteil des MedAT oberflächenähnlich sind.[8]

Aufnahmeverfahren für medizinische Studiengänge sind daher in Europa keine Seltenheit, sondern vielmehr der Regelfall. Hinsichtlich der Art der Aufnahmeverfahren und deren Einheitlichkeit innerhalb eines Landes bestehen jedoch große Unterschiede. Mit wenigen Ausnahmen (z.B. Italien, Schweiz, Vereinigtes Königreich) gibt es in den meisten anderen Ländern kein landesweit einheitliches, verbindliches Aufnahmeverfahren.

Im Vergleich zu Österreich ist auch interessant, dass viele Länder die bisherigen Noten der StudienplatzwerberInnen aus der Sekundarstufe bei der Aufnahme zum Medizinstudium miteinbeziehen. Generell zeigt eine Metaanalyse aus Deutschland, dass Schulnoten zwar valide Prädiktoren des späteren Studienerfolgs sind[9], jedoch weisen sie auch eine geringere Vergleichbarkeit und somit auch eine geringere Fairness als standardisierte Aufnahmeverfahren auf. Noten aus unterschiedlichen Regionen und Schulzweigen sind aktuellen empirischen Studien zufolge nur schwer direkt vergleichbar und korrelieren auch nur moderat mit landesweit durchgeführten standarisierten Schulleistungstests, mit deren Hilfe ein vergleichbares Stoffgebiet abgefragt wird.[10] Deshalb forderte das deutsche Bundesverfassungsgericht zurecht vom Gesetzgeber eine Angleichung der bundeslandspezifischen Noten.[11] Noten als Aufnahmekriterien werden zudem noch deutlich schwerer handhabbar, wenn nicht nur regionale Unterschiede innerhalb eines Landes in den realisierten Lehrplänen und Benotungsstandards auf einen Nenner gebracht werden müssen, sondern wenn Noten aus unterschiedlichen Ländern als Kriterium über die Aufnahme zum Medizinstudium (mit-)entscheiden sollen. Gemeinsam ist den meisten Ländern auch, dass die Aufnahme zum Medizinstudium zumeist hoch selektiv ist. In Österreich lag die Selektionsquote über verschiedene Jahre hinweg bei 10 bis 15 % der zum Test angemeldeten StudienplatzwerberInnen. Diese Selektionsquote ergibt sich ganz allgemein nicht aus der Strenge der Aufnahmeverfahren oder der aufnehmenden Medizinischen Universitäten, sondern ausschließlich aus dem Verhältnis an interessierten Studien-

7 McManus et al. (2013)
8 Lievens et al. (2016)
9 Haag et al. (2020)
10 Trapmann et al. (2007)
11 Trapmann et al. (2007)

platzwerberInnen zu den staatlich finanzierten und zur Verfügung gestellten Ausbildungsplätzen an den öffentlichen Medizinischen Universitäten in Österreich.[12]

# 3. Ausgangspunkt der Entwicklung des MedAT

An den Medizinischen Universitäten Wien, Graz und Innsbruck wurden, wie oben bereits erwähnt, ab dem Studienjahr 2006/07 Auswahlverfahren durchgeführt. An den Medizinischen Universitäten Wien und Innsbruck handelte es sich hierbei um den Schweizer Eignungstest für das Medizinstudium (EMS), während an der Medizinischen Universität Graz ein eigens entwickelter Kenntnistest (Basiskenntnistest Medizinische Studiengänge, BMS) eingesetzt wurde. Beide Aufnahmeverfahren wurden in einer unabhängigen Studie[13] hinsichtlich ihrer Anwendbarkeit in Österreich evaluiert. Die Ergebnisse der Evaluation zeigten, dass männliche Studienplatzwerber sowie StudienplatzwerberInnen aus Deutschland in beiden Aufnahmeverfahren höhere Punktewerte erzielten als deren Mitbewerberinnen, wobei ein Teil dieser Unterschiede auf eine systematische Bevorzugung der genannten Personengruppen zurückzuführen war. Dies bedeutete, dass selbst bei gleicher Fähigkeit männliche Studienplatzwerber bzw. StudienplatzwerberInnen aus Deutschland bei einigen Aufgaben eine höhere Lösungswahrscheinlichkeit hatten als ihre gleich fähigen österreichischen Mitbewerberinnen und zum Teil bereits durch ihre Gruppenzugehörigkeit höhere Punktewerte erzielten. Dies stellte ein Problem dar, da eine systematische Benachteiligung bzw. Bevorzugung bei gleicher Fähigkeit dem Prinzip der Gleichbehandlung widerspricht. Da Aufnahmeverfahren Wettbewerbssituationen darstellen, kommt dem Prinzip der Fairness und Gleichbehandlung im Rahmen eines Aufnahmeverfahrens eine ebenso zentrale Rolle zu wie der prognostischen Validität des Aufnahmeverfahrens.[14] Zwischen 2012 und 2013 wurden die Medizinischen Universitäten in Österreich im Rahmen ihrer Leistungsvereinbarungen daher beauftragt, ein gemeinsames österreichweites Aufnahmeverfahren für Human- und Zahnmedizin zu entwickeln, das aktuellen psychometrischen Standards standhält.

12 Neumann et al. (2009); Zimmermann et al. (2018)

13 Spiel et al. (2008)

14 zusammenfassend: APA, & NCME (2014); Arendasy et al. (2018); VfGH-Entscheidung VfSlg 19.899/2014

# 4. Beschreibung und Entwicklung des MedAT

Wie in der DIN 33430 (Qualitätsbezogene Anforderungen in der beruflichen Eignungsdiagnostik) empfohlen, begann der Prozess der Konstruktion des österreichweiten einheitlichen Aufnahmeverfahrens MedAT mit einer Anforderungsanalyse. Hierzu wurde von der Medizinischen Universität Wien eine Delphi-Studie durchgeführt, in deren Rahmen Lehrende befragt wurden, welche Fähigkeiten und Kenntnisse für einen erfolgreichen Abschluss des Studiums der Human- und Zahnmedizin erforderlich sind. Die Delphi-Studie wurde zudem auf der Grundlage von Befunden aktueller Metaanalysen aus dem deutschsprachigen und internationalen Raum zu Prädiktoren des Ausbildungserfolgs in medizinischen Studiengängen ergänzt.[15] Dass diese Befunde auch auf Österreich generalisierbar sind, zeigt eine aktuelle Evaluation der Zugangsregelungen zum Hochschulstudium in Österreich.[16] Auf Basis dieser Befunde wurden für das österreichweite Aufnah-

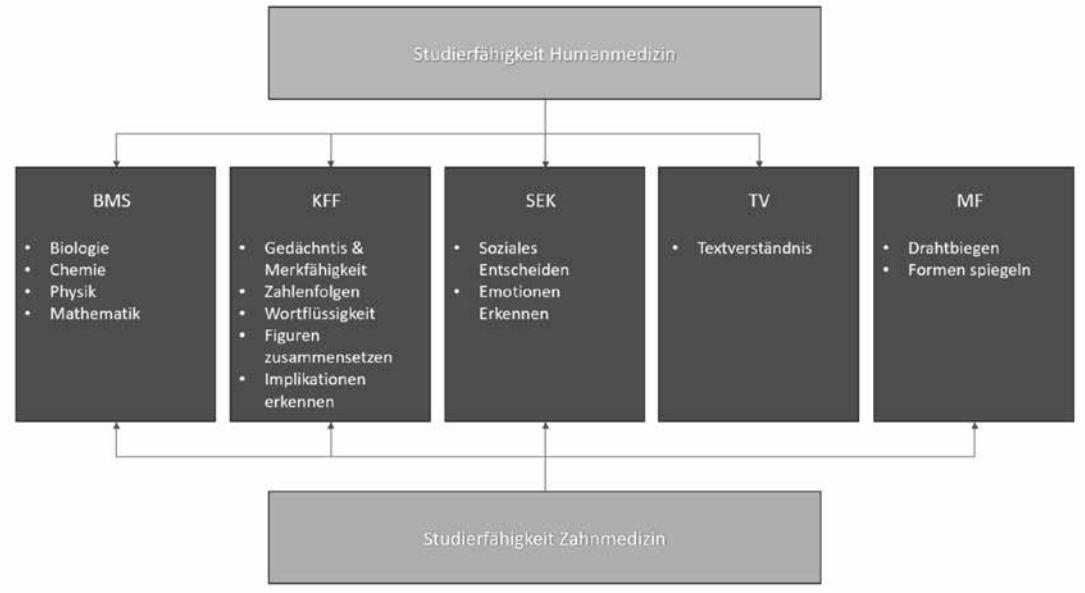


Abbildung 1: Testteile und Aufgabengruppen des MedAT

15 vgl. Donnon et al. (2007); Hell et al. (2007); Kuncel et al. (2010); Lievens (2004); McManus et al (2013); Schult et al. (2019)

16 Haag et al. (2020)

meverfahren Fähigkeits- und Kompetenzbereiche abgeleitet, die mit Hilfe des Aufnahmeverfahrens erfasst werden sollten. Die nachfolgende Abbildung gibt einen Überblick über die Fähigkeiten und Kenntnisse, die im Rahmen des aktuellen Aufnahmeverfahrens MedAT erfasst werden.

Der Testteil Basiskenntnistest Medizinische Studiengänge (BMS) ist hierbei inhaltlich weitgehend identisch mit dem früheren Aufnahmeverfahren der Medizinischen Universität Graz. Die Aufgaben dieses Testteils wurden auf Basis der bereits zuvor zitierten Evaluierung des früheren Aufnahmeverfahrens überarbeitet und optimiert. Dieser Testteil weist auch einige Gemeinsamkeiten mit dem Hamburger Naturwissenschaftstest (HAM-Nat) auf, der ebenfalls die Aufgabengruppen Biologie, Chemie, Physik und Mathematik erfasst. Sowohl der BMS als auch der HAM-Nat erwiesen sich in verschiedenen Studien als prädiktiv valide für die Vorhersage des Ausbildungserfolgs in der Human- und Zahnmedizin.[17] Im Gegensatz zum Testteil BMS umfasst der HAM-Nat darüber hinaus auch Aufgaben zum logischen Denken. Diese sind im MedAT im Testteil Kognitive Fähigkeiten und Fertigkeiten (KFF) enthalten. Konkret handelt es sich hierbei um die beiden Aufgabengruppen Zahlenfolgen und Implikationen erkennen. Der Testteil BMS blieb seit seiner ersten Anwendung im Jahr 2013 weitgehend unverändert, wobei jedoch jährlich aus Gründen der Testsicherheit und Fairness neu konstruierte Aufgaben vorgegeben werden.

Der Testteil Kognitive Fähigkeiten und Fertigkeiten (KFF) weist hinsichtlich seiner Aufgabenauswahl nicht nur Gemeinsamkeiten mit dem HAM-Nat auf, sondern auch mit dem Schweizer Eignungstest für das Medizinstudium (EMS) und dem Deutschen Test für Medizinische Studiengänge (TMS), die ebenfalls Aufgaben zur Raumvorstellung und zur Erfassung des Langzeitgedächtnisses enthalten. Alle drei genannten Verfahren erwiesen sich in bisherigen Studien als valide Prädiktoren des Studienerfolgs in der Human- und Zahnmedizin.[18] Bei der konkreten Wahl der Raumvorstellungsaufgaben unterscheiden sich jedoch die genannten Verfahren beträchtlich. Der Testteil KFF wurde erstmals im Jahr 2013 eingesetzt, wobei damals die Aufgabengruppen Implikationen Erkennen und Wortflüssigkeit noch nicht inkludiert waren. Diese sind erst im Jahr 2014 hinzugekommen. Ein weiterer Unterschied zwischen dem Testteil KFF und dem EMS bzw. dem TMS besteht darin, dass beim MedAT Textverständnis einen eigenen Testteil des Aufnahmeverfahrens für Humanmedizin bildet, während Aufgaben zum Textverständnis beim EMS und TMS in den kognitiven Teil integriert wurden.

17 vgl. Hissbach et al. (2011); Reibnegger et al. (2010, 2011)
18 vgl. Hell et al. (2007); Kraft et al. (2013); Trost et al (1997); Vetter & Sommer (2012)

Ein weiterer Unterschied zwischen MedAT und anderen deutschsprachigen Aufnahmeverfahren für Human- und Zahnmedizin zeigt sich im Testteil Soziale und Emotionale Kompetenzen (SEK), der in den Jahren 2015 und 2017 ergänzt wurde. In diesem Testteil werden Aufgaben zum sozialen Entscheidungsverhalten sowie zum Emotionalen Verständnis vorgegeben. Eine Erweiterung dieses Testteils um eine Skala zur Messung des Wissens über effektive Emotionsregulationsstrategien befindet sich aktuell in Entwicklung und wird derzeit noch evaluiert.[19] Äquivalente Testteile finden sich in der aktuellen Version des EMS und des TMS nicht, jedoch befindet sich eine Skala zur Erfassung einer der drei im MedAT berücksichtigten Facetten sozialer und emotionaler Kompetenzen in Deutschland aktuell in Entwicklung.[20] Der Grund für die Aufnahme dieses Testteils in das Aufnahmeverfahren MedAT bestand darin, dass neben dem Studienerfolg auch spätere Leistungen in Kursen mit kommunikativen und zwischenmenschlichen Inhalten sowie der spätere Umgang mit PatientInnen und KollegInnen prognostiziert werden sollen. Aktuelle Metaanalysen zeigen, dass soziale und emotionale Fähigkeiten zwar einen eher geringen Beitrag zur Prognose des Studienerfolgs leisten, jedoch inkrementell zur Prognose des Kommunikationsverhaltens mit KollegInnen und PatientInnen oder der Beurteilung der Leistung in der klinischen Praxis durch Ausbildende beitragen können.[21]

Neben den oben genannten Unterschieden in der Auswahl und Zusammenstellung der einzelnen Testteile und Aufgabengruppen unterscheidet sich der MedAT von anderen deutschsprachigen Aufnahmeverfahren für Human- und Zahnmedizin vor allem auch in der Art der Konstruktion des Testverfahrens. Während mit wenigen Ausnahmen[22] neben dem MedAT die meisten Aufnahmeverfahren für medizinische Studiengänge im deutschsprachigen Raum noch immer im Rahmen der Klassischen Testtheorie von menschlichen Aufgabenschreibern entwickelt werden, greift der MedAT auf einen anderen Ansatz zurück, der sich beispielsweise in den USA im Rahmen der Entwicklung des Scholastic-Aptitude-Tests (SAT) bereits empirisch gut bewährt hat. Bei der Entwicklung der Aufgabengruppen des MedAT werden neben Methoden der Klassischen Testtheorie auch Ansätze der Item-Response-Theorie (IRT) herangezogen. Der zentrale Vorteil der IRT besteht darin, dass sie ein formales mathematisches Modell über den Zusammenhang zwischen der Fähigkeit einer Person und ihrem Antwortverhalten bei der Bearbeitung einer Aufgabengruppe liefert. Die Passung dieses Modells auf die empirischen Daten kann empirisch evaluiert werden, was zugleich auch relevante

19 Arendasy et al. (2016c); Arendasy & Sommer (2022)
20 Schwibbe et al. (2018)
21 vgl. Cousans et al. (2017); Libbrecht & Lievens (2012); Libbrecht et al. (2014); MacCann et al. (2020); Patterson et al. (2016, 2017); Sànchez-Álvarez et al. (2020); Somaa et al. (2021)
22 Hissbach et al. (2011)

Informationen zu den Gütekriterien der betreffenden Aufgabengruppe liefert.[23] In diesem Sinne stellen Modelle der IRT eine ideale Ergänzung und Erweiterung der Methoden der Klassischen Testtheorie dar. Ein weiterer Unterschied zwischen dem MedAT und anderen Aufnahmeverfahren für medizinische Studiengänge im deutschsprachigen Raum besteht im Ansatz der Konstruktion der Aufgabengruppen. Im Rahmen der Entwicklung des MedAT wurde eine Kombination aus (1) einer theoriebasierten Aufgabenkonstruktion unter Zuhilfenahme menschlicher Aufgabenschreiber und (2) einer vollautomatischen Aufgabenkonstruktion mit Hilfe vorab empirisch evaluierter Aufgabengeneratoren angewandt.[24] Der letztgenannte Ansatz soll im nächsten Kapitel anhand eines Beispiels aus dem Testteil KFF kurz beschrieben werden.[25]

# 5. Prozess der Testkonstruktion im Rahmen der automatisierten Aufgabenkonstruktion

Bei der vollautomatischen Aufgabenkonstruktion (AIG) handelt es sich um einen Ansatz der Testkonstruktion, bei dem Aufgaben anhand einer systematischen Variation ihrer Aufgabenmerkmale so gestaltet werden, dass konstruktrelevante Denkprozesse durch sie angesprochen werden.[26] In einigen Ansätzen der automatischen Aufgabenkonstruktion, wie beispielsweise dem automatischen Min-Max-Ansatz, werden zudem auch Aufgabenmerkmale erarbeitet, die zu konstrukttirrelevanten Bearbeitungsstrategien und Denkprozessen einladen können. Diese Aufgabenmerkmale sollten bei den Aufgaben eines Aufnahmeverfahrens nicht vorkommen, da sie zu einer systematischen Bevorzugung oder Benachteiligung einzelner Gruppen an StudienplatzwerberInnen führen könnten.[27] Um die Denkprozesse und Bearbeitungsstrategien zu identifizieren, die für eine zu messende Fähigkeit charakteristisch sind (konstruktrelevante Denkprozesse), greifen Ansätze zur automatischen Aufgabenkonstruktion auf theoretische Modelle der Kogni-

23 Fischer & Molenaar (1995); Rost (2004) van der Linden & Hambleton (1997)
24 Arendasy & Sommer (2011); Arendasy & Sommer (2012); Irvine & Kyllonen (2002)
25 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)
26 Arendasy & Sommer (2011); Arendasy & Sommer (2012); Irvine & Kyllonen (2002)
27 Arendasy & Sommer (2011); Arendasy & Sommer (2012); Irvine & Kyllonen (2002)

tions- und Neurowissenschaften sowie empirische Befunde zu diesen Modellen zurück. Ähnlich wird auch bei der Spezifikation der Denk- und Bearbeitungsstrategien vorgegangen, die nicht für eine bestimmte Fähigkeit charakteristisch sind, durch die sich jedoch zumindest manche Aufgaben ebenfalls lösen lassen. In einigen Ansätzen der automatisierten Aufgabenkonstruktion, wie beispielsweise beim für Konstruktion der MedAT-Aufgaben angewandten automatisierten Min-Max-Ansatz, sollen solche Denk- und Bearbeitungsstrategien unterbunden werden, indem auf Aufgabenmerkmale verzichtet wird, die eine solche konstruktirrelevante Bearbeitungsweise ansprechen. Beim automatisierten Min-Max-Ansatz umfasst der Aufgabengenerator daher auch zwei Komponenten: eine generative Komponente mit deren Hilfe Aufgabenmerkmale systematisch variiert werden können, die konstruktrelevante Denkprozesse ansprechen, und eine Qualitätssicherungskomponente, durch die Aufgabenmerkmale unterdrückt werden, die zu einer alternativen Bearbeitung der Aufgaben einladen. Man spricht daher beim Min-Max-Ansatz auch von einem Zwei-Komponenten-Aufgabengenerator.[28] Der Prozess der Entwicklung und empirischen Evaluierung eines solchen Zwei-Komponenten-Aufgabengenerators ist in Abbildung 2 grafisch dargestellt.

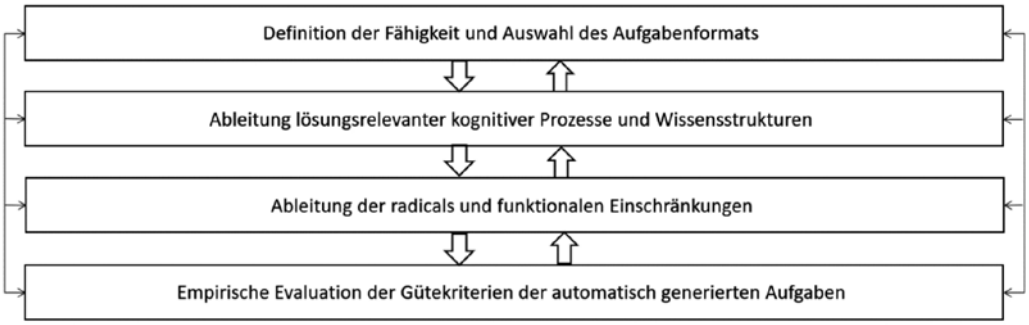


Abbildung 2: Prozessablauf der Testkonstruktion (modifiziert nach Arendasy & Sommer, 2011)

Ein Vorteil dieses Ansatzes besteht darin, dass sich hierdurch nachweislich die Validität und Fairness der so konstruierten Aufgaben deutlich verbessern lässt.[29] Darüber hinaus lässt sich auf diese Weise auch eine hohe Anzahl qualitativ hochwertiger Aufgaben konstruieren, die nicht nur den Bedarf an Aufgaben für die Aufnahmeverfahren selbst abdeckt, sondern auch den Bedarf an Aufgaben als Übungsmaterialen zur Vorbereitung auf das Aufnahmeverfahren, was die Fair-

28 Arendasy & Sommer (2011); Arendasy & Sommer (2012); Irvine & Kyllonen (2002)
29 Arendasy & Sommer (2011); Arendasy & Sommer (2012); Irvine & Kyllonen (2002)

ness im Sinne der Chancengleichheit verbessert.[30] Wie von verschiedener Seite oftmals betont, sind diese Anforderungen von menschlichen AufgabenschreiberInnen oft nur sehr schwer erfüllbar.[31] Ein weiterer Vorteil des hier verwendeten Min-Max-Ansatzes besteht zudem darin, dass mit Hilfe der Qualitätssicherheitskomponente auch die Eindeutigkeit der als richtig bewerteten Lösung formal-maschinell prüfbar werden kann.

## 5.1. Definition der Fähigkeitsdimension und Auswahl eines Aufgabenformats

Wie bereits zuvor beschrieben, wurden die zu messenden Fähigkeiten des MedAT aus aktuellen Metaanalysen und einer Delphi-Studie abgeleitet.[32] Beispielsweise zeigte sich hier, dass nicht nur logisches Denken, sondern auch Raumvorstellung für die Prognose des Studienerfolgs in Human- und Zahnmedizin relevant sind, und daher auch in aktuellen Aufnahmeverfahren im deutschsprachigen Raum erfasst werden. Bei der Wahl des Aufgabenformats zeigen sich jedoch Unterschiede zwischen den deutschsprachigen Aufnahmeverfahren. Im Rahmen des automatischen Min-Max-Ansatzes ist das Aufgabenformat so zu wählen, dass (1) konstruktrelevante Denkprozesse durch spezifische Gestaltungsmerkmale der Aufgaben angesprochen werden, während (2) konstruktirrelevante Denkprozesse, die zu einer Beeinträchtigung der Mess- oder Prognosefairness führen könnten, möglichst unterbunden werden sollten. Um dies zu bewerkstelligen, wurde im MedAT anstelle von Schlauchfiguren mit offenen Enden Aufgaben zum Zusammensetzen von Figuren ausgewählt, um Raumvorstellung zu erfassen. Das liegt darin, dass in verschiedenen Studien gezeigt werden konnte, dass Aufgaben zum Figurenzusammensetzen im Vergleich zu Schlauchfiguren eine geringere Wahrscheinlichkeit aufweisen, einzelne Gruppen an StudienplatzwerberInnen systematisch zu bevorzugen bzw. deren Studienerfolg systematisch zu unterschätzen.[33] Ähnliches gilt auch für die Wahl der beiden Aufgabenformate zur Messung des logischen Denkens, bei deren Auswahl ebenfalls Überlegungen zur Mess- und zur Prognosefairness unter Berücksichtigung von Gruppenunterschieden in den Testleistungen und späteren Studienleistungen eine zentrale Rolle spielten.[34]

30 zusammenfassend: Arendasy et al. (2018)
31 Hornke & Habon (1986)
32 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)
33 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)
34 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)

## 5.2. Ableitung lösungsrelevanter kognitiver Prozesse und Wissensstrukturen

In einem nächsten Schritt wird unter Rückgriff auf theoretische Modelle und empirische Befunde der Kognitions- und Neurowissenschaften definiert, welche Denkprozesse charakteristisch für die zu messende Fähigkeit sind. Diese sollen durch die Gestaltung der Aufgaben direkt angesprochen werden. Die so ermittelten Denkprozesse und Lösungsstrategien werden in ein einheitliches theoretisches Rahmenmodell integriert. Dieses bezieht sich noch nicht direkt auf einen bestimmten Aufgabentyp, sondern ist in dieser Phase noch auf verschiedene Aufgabentypen zur Messung der interessierenden Fähigkeit generalisierbar. Der direkte Bezug zu einem bestimmten Aufgabentyp wird erst im nächsten Entwicklungsschritt hergestellt.

## 5.3. Ableitung der Radicals und funktionalen Einschränkungen

In diesem nächsten Entwicklungsschritt wird das zuvor beschriebene theoretische Rahmenmodell für einen bestimmten Aufgabentyp ausgearbeitet. In dieser Phase werden nun auch bereits Aufgabenmerkmale betrachtet, die diese charakteristischen Denkprozesse ansprechen. Das zuvor beschriebene theoretische Rahmenmodell wird also zu einem Prozessmodell der Bearbeitung einer bestimmten Aufgabengruppe ausgearbeitet. Aus diesem Prozessmodell soll hervorgehen, wie die Testpersonen die Aufgaben konkret lösen und durch welche Aufgabenmerkmale diese Denkprozesse systematisch erschwert oder erleichtert werden.[35] Diese Aufgabenmerkmale werden in der Literatur als Radicals bezeichnet, da sie sich in vorhersagbarer Weise auf die Schwierigkeit der Aufgaben auswirken.[36] Die so ermittelten Radicals werden in die generative Komponente des Aufgabengenerators implementiert und bei der Konstruktion der Aufgaben systematisch variiert.

Darüber hinaus werden in Zwei-Komponenten-Ansätzen wie dem automatischen Min-Max-Ansatz auch Denkprozesse und Bearbeitungsstrategien definiert, die nicht erfasst werden sollen, da diese nicht charakteristisch für die zu messende Fähigkeit sind oder mit einem erhöhten Risiko einer systematischen Benachteiligung einer Gruppe an StudienplatzwerberInnen einhergehen. Aufgabenmerkmale, die zu einer solchen konstruktirrelevanten Bearbeitung einladen, werden in der

35 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)
36 Irvine (2002)

Literatur als funktionale Einschränkungen bezeichnet.[37] Sie werden in die Qualitätssicherungskomponente des Aufgabengenerators implementiert und sollen bei der Aufgabenkonstruktion unterdrückt werden, um die Fairness und Validität der automatisch konstruierten Aufgaben zu verbessern.

## 5.4. Empirische Evaluation der Gütekriterien der automatisch generierten Aufgaben

Wurde der Aufgabengenerator erst einmal entwickelt und implementiert, lässt sich mit dessen Hilfe eine beliebig hohe Anzahl an Aufgaben generieren, deren Gütekriterien in empirischen Studien evaluiert werden müssen. Alle im Rahmen des MedAT verwendeten Aufgabengeneratoren wurden bereits im Vorfeld der Entwicklung des MedAT konstruiert und empirisch evaluiert. Die Ergebnisse aus diesen Forschungsarbeiten wurden hierbei in Zeitschriften mit Peer-Review-Verfahren publiziert.[38] Die hierbei berücksichtigten Gütekriterien sollen aufgrund ihrer allgemeinen Relevanz für Aufnahmeverfahren im folgenden Abschnitt kurz dargestellt werden.

37 Greeno et al. (1993)
38 zusammenfassend: Arendasy, Sommer, & Feldhammer-Kahr (2020a)

# 6. Anforderungen an Gütekriterien von Aufnahmeverfahren

Da mit dem Ergebnis eines Aufnahmeverfahrens weitreichende Konsequenzen für den weiteren Lebensweg der StudienplatzwerberInnen verbunden sind, müssen Schlussfolgerungen, die aus den Ergebnissen eines Aufnahmeverfahrens gezogen werden, empirisch abgesichert werden. In der Literatur werden verschiedene Schlussfolgerungen und Annahmen beschrieben, die bei der Interpretation der Ergebnisse von Aufnahmeverfahren vorkommen, um eine Entscheidung über die Aufnahme der StudienplatzwerberInnen anhand von deren Testergebnissen treffen zu können. Diese Schlussfolgerungen und Annahmen korrespondieren mit verschiedenen Gütekriterien.[39] Der Zweck der Gütekriterien besteht darin, empirische Belege für die Gültigkeit aller Annahmen und Schlussfolgerungen vorzulegen, die im Rahmen der Entscheidung über die Aufnahme aus den Ergebnissen des Aufnahmeverfahrens abgeleitet werden müssen. Die hierbei erforderlichen empirischen Belege sind somit keineswegs untereinander austauschbar, bauen jedoch zumindest zum Teil logisch konzeptuell hierarchisch aufeinander auf (vgl. Abbildung 3).

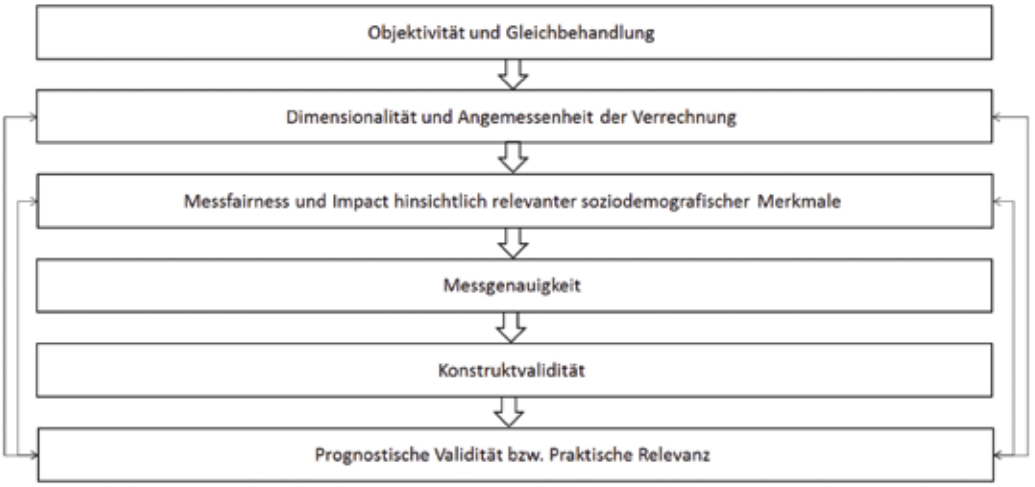


Abbildung 3: Anforderungen an die Gütekriterien eines Aufnahmeverfahrens

39 AERA, APA, & NCME (2014); Testkuratorium (2006, 2007, 2018)

# 7. Objektivität und Gleichbehandlung

Unter Objektivität versteht man das Ausmaß, in dem die Vorgabe, Auswertung und Interpretation der Ergebnisse unabhängig von den AuswerterInnen sind. In der Fachliteratur spricht man daher häufig von Durchführungs-, Auswertungs- und Interpretationsobjektivität.[40] Die Durchführungs- und Auswertungsobjektivität lassen sich durch einen hohen Grad an Standardisierung der Durchführung und Auswertung sicherstellen.[41] Um die Durchführungsobjektivität zu gewährleisten, gibt es zu allen Aufgabengruppen standardisierte Instruktionen und Anweisungen zur Vorgabe und zum Umgang mit Zwischenfragen. Die Zeitlimits der einzelnen Aufgabengruppen sind so gewählt, dass auch Personen mit Beeinträchtigungen innerhalb der vorgesehenen Zeit die Aufgaben bearbeiten können. Zudem werden für Personen mit Beeinträchtigungen spezielle Adaptierungen angeboten, die eine Vergleichbarkeit der Testbedingungen gewährleisten. Die Auswertung der Antworten erfolgt vollständig automatisiert nach einem eindeutigen Lösungsschlüssel, aus dem hervorgeht, welche der vorgegebenen Antwortmöglichkeiten als richtig zu verrechnen ist. Die Korrektheit dieses Lösungsschlüssels wird bei allen Aufgabengruppen vorab von menschlichen Aufgabenschreibern und gegebenenfalls von der Qualitätssicherungskomponente des entsprechenden Aufgabengenerators sichergestellt. Diese prüft hierbei, dass für jede einzelne Aufgabe ausschließlich die als richtig bewertete Antwortalternative als korrekt bewertet werden kann. Dies gewährleistet ein möglichst hohes Maß an Verrechnungssicherheit und Verrechnungseindeutigkeit. Zudem wird auf diese Weise auch die Gleichbehandlung aller StudienplatzwerberInnen bei der Durchführung und Auswertung des Aufnahmeverfahrens sichergestellt. Hierbei handelt es sich um eine zentrale Facette der Fairness, die konzeptuell eng mit dem Gütekriterium der Objektivität verbunden ist.[42] Konkret wird sichergestellt, dass alle StudienplatzwerberInnen unter vergleichbaren Bedingungen getestet werden und dass die Bewertung der Antworten aller StudienplatzwerberInnen nach einem eindeutigen, objektiven und transparenten Lösungsschlüssel erfolgt. Daher kann von einer Durchführungs- und Auswertungsobjektivität sowie von einer Gleichbehandlung aller StudienplatzwerberInnen ausgegangen werden. Die Interpretationsobjektivität wird durch klare Vorgaben sichergestellt, wie aus den Testergebnissen der einzelnen Aufgabengruppen Entscheidungen über die Aufnahme an einer der

40 Ziegler & Bühner (2012)
41 Kunnan (2000)
42 AERA, APA, & NCME (2014); Kunnan (2000); Xi (2010)

Medizinischen Universitäten abgeleitet werden sollen. Diese Vorgaben werden in den jährlichen Auswertungsverordnungen der Medizinischen Universitäten geregelt und sind österreichweit vereinheitlicht. Durch dieses Vorgehen kann auch ein hoher Grad an Interpretationsobjektivität attestiert werden.

# 8. Dimensionalität und Angemessenheit der Verrechnung

Ähnlich wie bei anderen deutschsprachigen Aufnahmeverfahren für Human- und Zahnmedizin wird auch im MedAT pro Aufgabengruppe die Anzahl der richtig gelösten Aufgaben als Maß für die zu erfassende Fähigkeit herangezogen. Diesem Vorgehen liegt die Annahme zu Grunde, dass interindividuelle Unterschiede im Antwortverhalten bei jeder Aufgabengruppe jeweils durch eine einzige zu messende Fähigkeit kausal verursacht werden und dass die Anzahl der gelösten Aufgaben ein akkurates Maß der interessierenden Fähigkeitsunterschiede darstellt. Verschiedene Studien zeigen, dass diese Annahme jedoch nur dann berechtigt ist, wenn für die betreffenden Aufgabengruppen die Passung des 1PL-Rasch-Modells empirisch belegt werden konnte.[43] Lässt sich dies nicht empirisch belegen, besteht laut vorliegenden Studien die Gefahr, dass durch die fälschliche Verwendung der Anzahl der gelösten Aufgaben als Maß für die zu messende Fähigkeit mehr leistungsschwächere Personen ausgewählt werden, als aufgrund ihrer Fähigkeit gerechtfertigt wäre.[44] Im Gegensatz zu alternativen Aufnahmeverfahren im deutschsprachigen Raum wie dem TMS oder dem EMS wird die Annahme der Eindimensionalität und Angemessenheit der Art der Verrechnung daher jährlich empirisch evaluiert. Die bisher vorliegenden Befunde sprechen für die Annahme, dass für alle Aufgabengruppen des MedAT eine gute Passung des 1PL-Rasch-Modells erzielt werden kann.[45] Inhaltlich bedeutet das, dass es durch die Verwendung der Anzahl gelöster Aufgaben beim MedAT zu keiner systematischen Bevorzugung oder Benachteiligung leistungsstärkerer StudienplatzwerberInnen kommt, sondern dass deren Fähigkeit akkurat und reliabel erfasst wird.

43 vgl. de Boeck & Wilson (2004); Fischer (1974); Rasch (1980); Rost (2004)
44 Borsboom (2006); Borsboom et al. (2008); Millsap (2011); Millsap & Kwok (2004)
45 Arendasy et al. (2013, 2014, 2015, 2016a, 2016b, 2017, 2018, 2019, 2020b, 2021)

# 9. Messfairness und Impact hinsichtlich relevanter soziodemografischer Merkmale

Da Aufnahmeverfahren Wettbewerbssituationen darstellen, kommt der Messfairness eine zentrale Bedeutung. Das Gütekriterium der Messfairness ist eng mit dem Gütekriterium der Eindimensionalität und Angemessenheit der Verrechnung verknüpft.[46] Messfairness bedeutet, dass Personen aus unterschiedlichen soziodemografischen Gruppen (z.B. Geschlecht, Nationalität, sozioökonomischer Status etc.) bei gleicher Fähigkeit die gleiche Wahrscheinlichkeit haben sollen, eine Aufgabe zu lösen bzw. einen bestimmten Punktewert zu erreichen.[47] Messfairness schließt jedoch nicht aus, dass sich soziodemografische Gruppen in ihren mittleren Testergebnissen voneinander unterscheiden. Die Frage von Gruppenunterschieden in der Testleistung ist daher in gewisser Weise unabhängig von der Frage der Messfairness. Messfairness stellt jedoch sicher, dass bestehende Unterschiede in der Testleistung zwischen soziodemografischen Gruppen als tatsächlich real bestehende Fähigkeitsunterschiede interpretiert werden können.[48] Diese werden in der Fachliteratur als Impact oder Gap bezeichnet. Analysen zur Messfairness sollten daher immer auch durch Analysen zu Gruppenunterschieden in den Testleistungen ergänzt werden, da sich selbst real bestehende Fähigkeitsunterschiede negativ auf die Prognosefairness eines Aufnahmeverfahrens auswirken können, sofern es in den Testergebnissen zu Mittelwertsunterschieden kommt, die nicht auch im vergleichbaren Maß für die vorherzusagenden Kriteriumsmaße gelten. Eine Analyse der Messfairness ist somit unabhängig davon unabdingbar, ob sich zwei Gruppen an StudienplatzwerberInnen in ihren Testleistungen voneinander unterscheiden. Die zentrale Bedeutung der Messfairness von Aufnahmeverfahren zeigt sich auch in verschiedenen Studien, in denen nachgewiesen werden konnte, dass bei einer Verletzung der Messfairness, unabhängig von Gruppenunterschieden in der Testleistung, mehr Personen aus der systematisch bevorzugten Gruppe aufgenommen werden, als aufgrund ihrer tatsächlichen Eignung gerechtfertigt wäre.[49] Aktuelle Befunde zur Messfairness der Aufgabengruppen des MedAT hinsichtlich der Personenmerkmale Geschlecht,

46 Xi (2010); Kane (2010); Kunnan (2000)
47 Arendasy et al. (2018); Borsboom et al. (2008); Kunnan (2000); Millsap (2011); Mislevy et al. (2013)
48 vgl. Chen (2008); Li & Zumbo (2009)
49 Borsboom et al. (2008); Millsap & Kwon (2004)

Nationalität und sozioökonomischer Status sprechen für die Annahme, dass durch die Aufgaben des MedAT keine der genannten Personengruppen systematisch bevorzugt oder benachteiligt wird.[50] In ergänzenden Studien zur Messfairness konnte darüber hinaus auch noch gezeigt werden, dass Messfairness auch für weitere Personenmerkmale wie Muttersprache der StudienplatzwerberInnen, Standort der aufnehmenden Medizinischen Universität, Schultyp, wiederholte Antritte zum Aufnahmeverfahren, Art der Vorbereitung auf das Aufnahmeverfahren, Testangst und emotionale Bewertung der Situation der Aufnahmetestung angenommen werden kann.[51] Hierin liegt ein weiterer großer Unterschied zwischen dem MedAT und alternativen Aufnahmeverfahren für Human- und Zahnmedizin im deutschsprachigen Raum, in denen Analysen der Messfairness ausgespart werden, nur dann durchgeführt werden, wenn sich in einer Aufgabengruppe Mittelwertsunterschiede zwischen den interessierenden Gruppen zeigen, und in denen die Messfairness hinsichtlich relevanter Personenmerkmale zumindest für Österreich empirisch widerlegt wurde.[52]

50 Arendasy et al. (2013, 2014, 2015, 2016, 2016a, 2016b, 2017, 2018, 2019, 2020b, 2021)

51 vgl. Arendasy et al. (2016, 2016b); Sommer & Arendasy (2015, 2016); Sommer et al. (2019)

52 Hänsgen & Spicher (2013, 2014, 2015, 2016, 2017); Spicher (2018, 2019, 2020, 2021); Trost et al. (1998); Spiel et al (2008)

# 10. Messgenauigkeit

Da mit der Entscheidung über die Aufnahme der StudienplatzwerberInnen weitreichende Konsequenzen verbunden sind, muss ein Aufnahmeverfahren auch bestimmten Mindeststandards hinsichtlich seiner Messgenauigkeit genügen.[53] Beim MedAT basiert die Entscheidung über die Aufnahme zum Studium auf einem optimal gewichteten Summenwert aus allen Testteilen, der die Grundlage für die Rangreihung der StudienplatzwerberInnen bildet. Dieser sollte daher auch eine möglichst hohe Messgenauigkeit aufweisen. Die Messgenauigkeit kann mit Hilfe unterschiedlicher Maßen der inneren Konsistenz bestimmt werden, deren Werte zwischen 0 und 1 liegen, wobei bei Aufnahmeverfahren eine Messgenauigkeit von ≥ 0,90 erzielt werden sollte.[54] In den bisherigen Analysen des MedAT wurde die gewünschte Messgenauigkeit immer eingehalten bzw. sogar übertroffen. Über die verschiedenen Jahre hinweg schwankte die Messgenauigkeit zwischen 0,91 und 0,98 und erfüllt somit die Anforderungen an die Messgenauigkeit eines Aufnahmeverfahrens.[55] Die Messgenauigkeit des MedAT liegt somit nicht nur leicht oberhalb der Anforderungen an Aufnahmeverfahren, sondern bewegt sich auch in einer Größenordnung, die jener alternativer Aufnahmeverfahren zum Studium der Human- und Zahnmedizin im deutschsprachigen Raum entspricht.[56]

53 AERA, APA, & NCME (2014); Testkuratorium (2006, 2007, 2018)
54 Ziegler & Bühner (2012)
55 Arendasy et al. (2013, 2014, 2015, 2016a, 2016b, 2017, 2018, 2019, 2020b, 2021)
56 Hänsgen & Spicher (2013, 2014, 2015, 2016, 2017); Hissbach et al. (2011); Spicher (2018, 2019, 2020, 2021); Trost et al. (1998); Spiel et al. (2008)

# 11. Konstruktvalidität des MedAT

Bei der Konstruktion eines Aufnahmeverfahrens wurden Aufgabengruppen ausgewählt, die bestimmte relevante Fähigkeiten erfassen sollen. Bei der Überprüfung der Konstruktvalidität geht es um einen empirischen Nachweis, dass mit den ausgewählten Aufgaben tatsächlich die gewünschte Fähigkeit erfasst wird.[57] In der Forschung werden häufig zwei Facetten der Konstruktvalidität unterschieden: (1) die inhaltlich-logische Validität und Konstruktrepräsentation sowie (2) die Faktorenstruktur, konvergente und diskriminante Validität.

Die inhaltlich-logische Validität und die Konstruktrepräsentation befassen sich mit den Inhalten und den Denkprozessen, die bei der Bearbeitung der einzelnen Aufgabengruppen eine Rolle spielen. Die Schritte bei der Überprüfung der inhaltlich-logischen Validität und Konstruktrepräsentation sind bei Tests zur Messung kognitiver Fähigkeiten und studienfachrelevanter Vorkenntnisse ähnlich. Während bei Wissenstests die inhaltlich-logische Validität und Konstruktrepräsentation in der Regel mit Hilfe unterschiedlicher Expertinnen- und Experten-Panel-Methoden überprüft wird, liegt der Fokus bei kognitiven Tests auf Befunden aus den Neuro- und Kognitionswissenschaften mit deren Hilfe Aufgabenmerkmale erarbeitet werden, die die theoretisch postulierten Denkprozesse aktiv ansprechen sollen.[58] Dies bedeutet auch, dass die Schwierigkeit einer Aufgabe anhand von Aufgabenmerkmalen erklär- bzw. vorhersagbar sein sollte, durch die diese Denkprozesse angesprochen werden. Bei der empirischen Überprüfung der Konstruktrepräsentation geht es also darum herauszufinden, wie gut die theoretisch prognostizierten und die tatsächlich ermittelten Aufgabenschwierigkeiten übereinstimmen.[59] Für die Aufgabengruppen des Testteils KFF konnte gezeigt werden, dass die theoretisch prognostizierten und die tatsächlich ermittelten Aufgabenschwierigkeiten sehr hoch miteinander korrelieren (R = 0,88 bis 0,954). Dies bedeutet, dass 77,1 % bis 91,0 % der Unterschiede in den Testergebnissen der StudienplatzwerberInnen durch Unterschiede in der Beherrschung der als relevant erachteten Denkprozesse erklärt werden können.[60] Dieser Befund konnte auch anhand der Aufgaben-Sets aus den Jahren 2013 bis 2021 repliziert werden und kann daher als sehr stabil bezeichnet werden.

57 Embretson (1983); Kane (1992, 2001, 2010, 2016); Messick (1989, 1995)
58 vgl. Arendasy & Sommer (2011, 2012); Arendasy et al. (2020a)
59 vgl. Arendasy & Sommer (2011, 2012); Arendasy et al. (2020a)
60 für Details: Arendasy et al. (2020a)

Ergänzend hierzu befassen sich die konvergente und diskriminante Validität mit empirischen Nachweisen, dass (1) ein bestimmtes Aufnahmeverfahren die theoretisch erwartete Faktorenstruktur aufweist und (2) in theoretisch zu erwartender Weise mit konstruktnahen und konstruktfernen Tests zusammenhängt.[61] Bereits im Vorfeld der Entwicklung des MedAT wurden Studien zur Faktorenstruktur und zur konvergenten und diskriminanten Validität einzelner Aufgabengruppen durchgeführt und in Zeitschriften mit einem Peer-Review-Verfahren publiziert.[62] In diesen Studien konnte gezeigt werden, dass die Aufgaben in theoretisch zu erwartender Weise mit anderen Tests zusammenhängen. Darüber hinaus konnte in den folgenden Jahren bei der Analyse der Daten des MedAT auch gezeigt werden, dass mit den Aufgabengruppen tatsächlich die in Abbildung 1 postulierten vier Kompetenzen erfasst werden.[63] Abbildung 4 veranschaulicht die Ergebnisse zur Faktorenstruktur des MedAT.

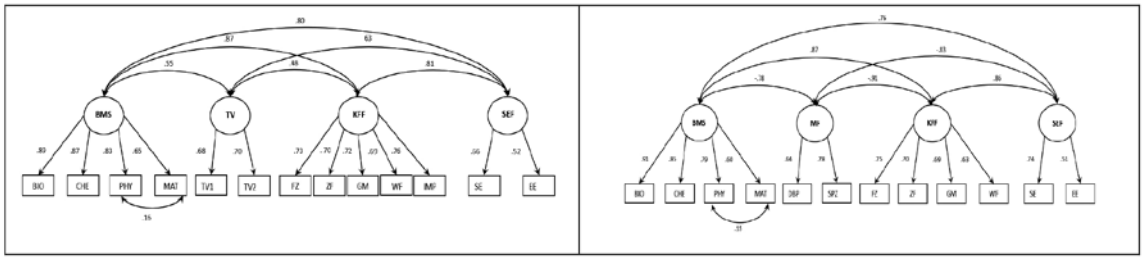


Abbildung 4: Faktorenstruktur des MedAT-H (links) und MedAT-Z (rechts)

Wie aus der Grafik ersichtlich, laden die Faktoren mit 0,51 bis 0,91 hoch auf die ihnen zugehörigen Aufgabengruppen. Zudem zeigte sich auch, dass die vier Faktoren, die mit den jeweils vier Testteilen korrespondieren, zwar hoch miteinander korrelieren, jedoch zugleich auch klar konzeptuell und empirisch voneinander unterscheidbar sind. Die Höhe der Korrelationen zwischen den vier Testteilen entspricht zudem den Erwartungen aufgrund der aktuellen Forschungsliteratur.54 Die vorliegenden empirischen Befunde sprechen somit zusammenfassend betrachtet für die Konstruktvalidität des Aufnahmeverfahrens MedAT und können aufgrund ihrer mehrfachen Replikation als sehr stabil beurteilt werden.

61 Embretson (1983); Kane (1992, 2001, 2010, 2016); Messick (1989, 1995)
62 zusammenfassend: Arendasy et al. (2020a)
63 vgl. Arendasy et al. (2013, 2014, 2015, 2016a, 2016b, 2017, 2018, 2019, 2020b, 2021)

# 12. Prognostische Validität

Im Rahmen der Beschreibung der Entwicklung des MedAT wurde bereits auf einige der aktuell vorliegenden Studien zur prognostischen Validität für die Vorhersage des Studienerfolgs und der Studiendauer eingegangen. Die vorliegenden deutschsprachigen und internationalen Studien und Metaanalysen zeigen, dass kognitive Fähigkeiten (wie im Testteil KFF des MedAT erfasst) prognostisch valide für die Vorhersage des späteren Studienerfolgs sind.[64] Ähnliches gilt auch für die naturwissenschaftlichen Grundkenntnisse, manuellen Fertigkeiten und Textverständnis.[65] In diesen Studien zeigte sich auch, dass naturwissenschaftliche Vorkenntnisse in den ersten Studienjahren eine höhere prognostische Validität haben als kognitive Fähigkeiten. Im Gegensatz dazu tragen kognitive Fähigkeiten stärker zur Vorhersage des Studienerfolgs in späteren Abschnitten bei. Dieser Befund konnte auch für eine Vorversion der KFF Aufgabengruppen des MedAT repliziert werden, in der sich zudem auch die prognostische Validität dieser Aufgabengruppe zur Vorhersage des United State Medical Licensing Examination (USMLE) zeigte.[66]

Weitere Hinweise auf die prognostische Validität des MedAT ergeben sich aus einer aktuellen Evaluation der Zugangsregelungen zum Hochschulstudium in Österreich.[67] In dieser Studie konnte zumindest für die ersten mit Hilfe des MedAT aufgenommenen Kohorten gezeigt werden, dass die Erfolgsquoten der ausgewählten Studierenden ähnlich bzw. zum Teil leicht höher ausfielen als in den Vorjahren, in denen die österreichischen Medizinischen Universitäten auf alternative Auswahlverfahren (konkret EMS oder BMS) zurückgriffen. Dieser Befund ist insofern bemerkenswert, als sich kurz nach Einführung des MedAT laut vorliegender Evaluation der Anteil der aufgenommenen und erfolgreich studierenden Frauen aus Österreich erhöhte, ohne dass es zu einer Änderung der Erfolgsquote kam.[68] Eine mögliche Erklärung für diesen Befund bestünde in dem zuvor beschriebenen Konstruktionsprinzip der Aufgabengruppen des MedAT, das eine systematische Benachteiligung und Unterschätzung der späteren Studienleistung österreichischer Studienplatzwerberinnen minimiert. Studien zur prognostischen Validität des Testteils soziale und emotionale Kompetenzen sowie Studien zur prognosti-

64 z.B.: Donnon et al. (2007); Hell et al. (2007); Kuncel et al. (2010); Lievens (2004); Lievens (2004); McManus et al (2013); Schult et al. (2019)
65 z.B.: Donnon et al. (2007); Reibnegger et al. (2010, 2011); Schult et al. (2019)
66 Vetter & Sommer (2012)
67 Haag et al. (2020)
68 z.B.: Donnon et al. (2007); Hell et al. (2007); Kuncel et al. (2010); Lievens (2004); Lievens (2004); McManus et al (2013); Schult et al. (2019)

schen Validität in Hinblick auf alle Testteile unter Berücksichtigung eines breiteren Spektrums relevanter Erfolgskriterien (z.B. Kommunikationsverhalten, Umgang mit PatientInnen) befinden sich aktuell noch in Arbeit.

Die bislang bereits vorliegenden Befunde sprechen jedoch für die Annahme, dass mit dem MedAT eine valide Prognose des späteren Studienerfolgs möglich ist. Zudem legen die vorliegenden Befunde auch nahe, dass sich die Prognosefairness des Aufnahmeverfahrens seit Einführung des MedAT hinsichtlich des soziodemografischen Merkmals Geschlecht verbessert hat, da es bei gleichbleibender Erfolgsquote und Studiendauer zu einem erhöhten prozentuellen Anteil aufgenommener weiblicher Studierender aus Österreich kam.[69] Dieser durchaus erfreuliche Befund kann möglicherweise auf weitere Merkmale von Gruppenzugehörigkeit (Nationalität, Vorbildung und sozioökonomischer Status) erweitert werden und soll in der sich in Arbeit befindlichen Validierungsstudie zum MedAT anhand der Daten mehrerer Kohorten an StudienplatzwerberInnen nochmals validiert bzw. kreuzvalidiert werden.

69 Haag et al. (2020)

# 13. Fazit zur Nützlichkeit von Aufnahmeverfahren wie dem MedAT

Um ein Aufnahmeverfahren als nützlich bezeichnen zu können, muss es einen oder mehrere Zwecke erfüllen, die durch kein alternatives Verfahren zur Aufnahme der StudienplatzwerberInnen gleich gut oder besser erfüllt werden können.[70]

Da im Studium der Human- und Zahnmedizin die Zahl der StudienplatzwerberInnen seit geraumer Zeit enorm gestiegen ist und mittlerweile das 10fache der Studienplatzzahl beträgt, ist es erforderlich, unter Wahrung der Qualität der Ausbildung die verfügbaren Kapazitäten möglichst effektiv zu nutzen. Hierzu sollen mit Hilfe eines Aufnahmeverfahrens jene StudienplatzwerberInnen einen Ausbildungsplatz erhalten, von denen angenommen werden kann, dass sie das Studium der Human- bzw. Zahnmedizin erfolgreich in möglichst kurzer Zeit absolvieren können. Hinsichtlich der Prognose des Studienerfolgs erwies sich MedAT ähnlich effektiv bzw. etwas effektiver als andere deutschsprachige Aufnahmeverfahren. Der wesentliche Vorteil des MedAT besteht vor allem in einer deutlichen Verbesserung der Fairness und Chancengleichheit für unterschiedliche Gruppen an StudienplatzwerberInnen. Dies ist nicht nur durch die Art der Konstruktion des MedAT unter Zuhilfenahme von Ansätzen der vollautomatischen Aufgabenkonstruktion zu erklären, sondern auch damit, dass der MedAT im Vergleich zu anderen deutschsprachigen sowie internationalen Aufnahmeverfahren ein breiteres Spektrum an relevanten Determinanten des Studien- und Berufserfolgs abdeckt. Durch MedAT-H und MedAT-Z kommt es damit zu keiner systematischen Bevorzugung oder Benachteiligung einzelner soziodemografischer Gruppen an StudienplatzwerberInnen, sodass zwei Personen mit gleicher Befähigung unabhängig von anderen Personenmerkmalen die gleiche Chance haben, einen bestimmten Testwert zu erzielen und aufgenommen zu werden. *Für die Wirkung von Aufnahmeverfahren werden* drei Kennzahlen angeführt: Anteil der prüfungsaktiven Studien (Studierenden), Anzahl der Studienabschlüsse an Universitäten innerhalb der Toleranzstudiendauer und die durchschnittliche Studiendauer.[71] Es wird als plausibel bezeichnet, dass die mit dem Zugangsverfahren (MedAT) ausgewählten Studierenden ein höheres Maß an Eignung aufwiesen und die Kennzahlen sich in

70 Ziegler & Bühner (2012)

71 Bericht des Rechnungshofes, Aufnahmeverfahren Human- und Zahnmedizin, Reihe BUND 2020/47; Haag et al (2020)

den letzten Jahren verbesserten. Zudem wird angeführt, dass auch andere Faktoren dazu beitragen, wie zum Beispiel die Curricula selbst. Die durchschnittlichen Studiendauern an den öffentlichen Medizinischen Universitäten liegen deutlich unter denen anderer Universitäten (alle Studien).[72]

Ein wesentlicher Vorteil standardisierter Aufnahmeverfahren wie MedAT, EMS, TMS, HAM-Nat etc. gegenüber alternativen Aufnahmekriterien wie Noten, Bewerbungsgespräche und der Erfassung von Praktika mittels Zeugnissen oder biografischen Inventaren besteht in ihren höheren Gütekriterien vor allem hinsichtlich Objektivität, Reliabilität, Validität und Fairness.[73] Darüber hinaus sind sie auch weniger einfach zu verfälschen, was im Kontext einer hochselektiven Wettbewerbssituation wie der Aufnahme zum Medizinstudium relevant ist, da es durch individuelle Unterschiede in der Bereitschaft zu verfälschen zu nennenswerten Unterschieden in den Rangreihungen und deren prädiktiven Validität kommen kann.[74]

Der Vorteil von fachspezifischen standardisierten Aufnahmeverfahren im Vergleich zu alternativen Aufnahmeverfahren zeigt sich auch in der Bewertung durch StudienplatzwerberInnen. So konnte beispielsweise in einer Studie aus Deutschland gezeigt werden, dass StudienplatzwerberInnen die Akzeptanz von fachspezifischen Aufnahmeverfahren höher bewerten als jene von Interviews, Praktika, Assessment Center, bisherigen Schulnoten, Motivationsschreiben, Wartelisten oder Losverfahren.[75]

72 Bericht des Rechnungshofes, Aufnahmeverfahren Human- und Zahnmedizin, Reihe BUND 2020/47; Haag et al (2020)
73 Bericht des Rechnungshofes, Aufnahmeverfahren Human- und Zahnmedizin, Reihe BUND 2020/47; Haag et al (2020)
74 Jäger (2003)
75 Stegt et al. (2018)

# Literatur

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: Author.

Arendasy, M., & Sommer, M. (2022). Erweiterung der Aufgabengruppe Soziale und emotionale Kompetenzen: Konzept für die Skala Emotionen Regulieren. Universität Graz.

Arendasy, M., & Sommer, M. (2011). Automatisierte Itemgenerierung: Aktuelle Ansätze, Anwendungen und Forschungen. In L. F. Hornke, M. Amelang, & M. Kersting (Hrsg.), Enzyklopädie für Psychologie: Methoden der Psychologischen Diagnostik (S. 215-280). Hogrefe.

Arendasy, M., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes assessment. Learning and Individual Differences, 22, 112–117.

Arendasy, M., Sommer, M., & Feldhammer, M. (2013). MedAT-H & MedAT-T 2013: Auswertungsdokumentation und psychometrische Evaluation. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer, M. (2014). MedAT-H & MedAT-T 2014: Auswertungsdokumentation und psychometrische Evaluation. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer, M. (2015). MedAT-H & MedAT-T 2015: Auswertungsdokumentation und psychometrische Evaluation. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer, M. (2016a). MedAT-H & MedAT-T 2016: Auswertungsdokumentation und psychometrische Evaluation. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer, M. (2016b). MedAT-H Wien: Ergänzende Auswertung zum sozio-ökonomischen Status, Schultyp, wiederholten Testantritt, und zur Muttersprache. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer, M. (2016b). Emotionen Erkennen und Soziales Entscheiden: Kurzfassung der theoretischen Grundlagen und deren Umsetzung. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2017). MedAT-H & MedAT-Z 2017: Psychometrische Evaluation. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2018). MedAT-H & MedAT-Z 2018: Psychometrische Evaluation. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2019). MedAT-H & MedAT-Z 2019: Psychometrische Evaluation. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2020a). Konstruktion KFF für MedAT: Dokumentation des Konstruktionsrationals. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2020b). MedAT-H & MedAT-Z 2020: Psychometrische Evaluation. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., & Feldhammer-Kahr, M. (2021). MedAT-H & MedAT-Z 2020: Psychometrische Evaluation. Technischer Bericht AB Psychologische Diagnostik & Methodik. Universität Graz.

Arendasy, M., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? Intelligence, 55, 44–56.

Arendasy, M., Sommer, M., Feldhammer-Kahr, M., Freudenthaler, H. H., Punter, F. J., Rieder, A. (2018). Fairness als zentrale Herausforderung moderner Aufnahmeverfahren. Zeitschrift für Hochschulentwicklung, 13, 37–54.

Arendasy, M., Sommer, M., Punter, F. J., Feldhammer-Kahr, M., & Rieder, A. (2019). Do individual differences in test-takers' appraisal of admission testing compromise measurement fairness? *Intelligence*, 73, 16–29.

Borsboom, D. (2006). When does measurement invariance matter? Medical Care, 44, 176–181.

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? Psychological Methods, 13, 75–98.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inap-propriate comparisons in cross-cultural research. Journal of Personality and Social Psychology, 95, 1005–1018.

Cousans, F., Patterson, F., Edwards, H., Walker, K., McLachlan, J. C., & Good, D. (2017). Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. Advances in Health Sciences Education, 22, 401–413.

de Boeck, P., & Wilson, M. (2004). Explanatory item response models. Springer New York.

Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. Academic Medicine, 82, 100–106.

Ebach, J. & Trost, G. (1997). Admission to Medical Schools in Europe. Berlin: Pabst Science Publishers

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179–197.

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

Fischer, G. H., & Molenaar, I. W. (1995). Rasch Models. Foundations, Recent Developments, and Applications. Berlin: Springer.

Greeno, J. G., Moore, J. L. & Smith, D. R. (1993). Transfer of situated learning. In Dettermann, D. K. & Sternberg, R. J. (Eds.). Transfer on trial: Intelligence, cognition, and instruction. (pp. 99–167). Wetsport, CT, US: Ablex Publishing.

Grimm, M., & Marschall, D. (2016). Zulassungsverfahren an Universitäten und Legitimation von Kostenbeiträgen. In Hauser, W. (Hrsg.). Hochschulrecht – Jahresbuch 2016. (S. 214–233). Wien, Graz: NWV.

Haag, N., Thaler, B., Stieger, A., Unger, M., Humpl, S. & Mathä, P. (2020). Evaluierung der Zugangsregelungen nach § 71b, § 71c, § 71d UG 2002. Wien: IHS.

Hampe, W. und Kadmon, M. (2019). Who is allowed to study medicine? regulations and evidence. GMS J Med Educ. 36(1). Doc10. doi: 10.3205/zma001218

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

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

Hänsgen K. & Spicher B. (2013). EMS – Eignungstest für das Medizinstudium in der Schweiz 2013: Bericht 20 über die Durchführung und Ergebnisse 2013. Fribourg: University of Fribourg.

Hänsgen K. & Spicher B. (2014). EMS – Eignungstest für das Medizinstudium in der Schweiz 2014: Bericht 21 über die Durchführung und Ergebnisse 2014. Fribourg: University of Fribourg.

Hänsgen K. & Spicher B. (2015). EMS – Eignungstest für das Medizinstudium in der Schweiz 2015: Bericht 22 über die Durchführung und Ergebnisse 2015. Fribourg: University of Fribourg.

Hänsgen K. & Spicher B. (2016). EMS – Eignungstest für das Medizinstudium in der Schweiz 2016: Bericht 23 über die Durchführung und Ergebnisse 2016. Fribourg: University of Fribourg.

Hänsgen K. & Spicher B. (2017). EMS – Eignungstest für das Medizinstudium in der Schweiz 2017: Bericht 24 über die Durchführung und Ergebnisse 2017. Fribourg: University of Fribourg.

Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. Empirische Pädagogik, 21, 251–270.

Hissbach, J. C., Klusmann, D., & Hampe, W. (2011). Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. BMC medical education, 11, 1–11.

Hornke, L. F. & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. Applied Psychological Measurement, 10, 369–380.

Irvine, S. H. (2002). The foundations of Item Generation for Mass Testing. In S.H. Irvine & P.C. Kyllonnen (Eds.). Item Generation for Test Development (p. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine, S. H. & Kyllonen, P. C. (2002). Item generation for test development. Mahwah, NJ: Lawrence Erlbaum.

Jäger, R. S. (2003). Biographisches Inventar. In K. D. Kubinger & R. S. Jäger (Hrsg.), Schlüsselbegriffe der Psychologischen Diagnostik (S. 61–67). Weinheim: Beltz.

Kane, M. (1992). An argument-based approach to validation. Psychological Bulletin, 112, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319–342.

Kane, M. (2010). Validity and Fairness. Language Testing, 27, 177–182.

Kane, M. T. (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice, 23, 198–211.

Kraft, H. G., Lamina, C., Kluckner, T., Wild, C., & Prodinger, W. M. (2013). Paradise lost or paradise regained? Changes in admission system affect academic performance and drop-out rates of medical students. Medical Teacher, 35, 1123–1129.

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Exami-nation for master's and doctoral programs: A meta-analytic investigation. Educational and Psychological Measurement, 70, 340–352.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), Fairness and validation in language assessment (pp. 1–14). Cambridge University Press.

Li, Z., & Zumbo, B.D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. Psicologica, 30, 343–370.

Libbrecht, N., & Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. International Journal of Psychology, 47, 438–447.

Libbrecht, N., Lievens, F., Carette, B., & Côté, S. (2014). Emotional intelligence predicts success in medical school. Emotion, 14, 64–73.

Lievens, F. (2004). Longitudinal study of the validity of different cognitive ability tests in a student admission context. Applied H.R.M Research, 9, 27–30.

Lievens, F., Patterson, F., Corstjens, J., Martin, S. und Nicholson, S. (2016). Widening access in selection using situational judgement tests: evidence from the UKCAT. Medical Education, 50, 624–636.

MacCann, C., Jiang, Y., Brown, L. E., Double, K. S., Bucich, M., & Minbashian, A. (2020). Emotional intelligence predicts academic performance: A meta-analysis. Psychological Bulletin, 146, 150–186.

McManus, I. C., Dewberry, C., Nicholson, S., Dowell, J. S., Woolf, K., und Potts, H. W. W. (2013). Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies. BMC Medicine, 11, 243–263.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: American Council on Education & Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' re-sponses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741–749.

Millsap, R. E. (2011). Statistical approaches to measurement invariance. Routledge.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. Psychological Methods, 9, 93–115.

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. Educational Research and Evaluation: An International Journal on Theory and Practice, 19, 121–140.

Neumann, M., Nagy, G., Trautwein, U., & Lüdtke, O. (2009). Vergleichbarkeit von Abiturleistungen. Zeitschrift für Erziehungswissenschaft, 12, 691–714.

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016). The predictive validity of a situational judgement test and multiple-mini-interview for entry into postgraduate training in Australia. BMC Medical Education, 16: 87.

Patterson, F., Cousans, F., Edwards, H., Rosselli, A., Nicholson, S., & Wright, B. (2017). The predictive validity of a text-based Situational Judgment Test in undergraduate medical and dental school admissions. Academic Medicine, 92, 1250–1253.

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

Reibnegger, G., Caluba, H.-C., Ithaler, D., Manhal, S., Neges, H. M., & Smolle, J. (2010). Progress of medical students after open admission or admission based on knowledge tests. Medical Education, 44, 205–214.

Reibnegger, G., Caluba, H.-C., Ithaler, D., Manhal, S., Neges, H. M., & Smolle, J. (2011). Dropout rates in medical students at one school before and after the installation of admission tests in Austria. Academic Medicine, 86, 1040–1048.

Rost, J. (2004). Lehrbuch Testtheorie, Testkonstruktion. Huber.

Sànchez-Álvarez, N., Mortes, M. P. B., & Extremera, N. (2020). A meta-analysis of the relationship between emotional intelligence and academic performance in secondary education: A multi-stream comparison. Frontiers in Psychology, 11, 1–11.

Schult, J., Hofmann, A. & Stegt, S. J. (2019). Leisten fachspezifische Studierfähigkeitstests im deutschsprachigen Raum eine valide Studienerfolgsprognose? Ein metaanalytisches Update. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 51, 1–15.

Schwibbe, A., Kackamp, J., Knorr, M., Hissbach, J., Kadmon, M., und Hampe, W. (2018). Medizin-studierendenauswahl in Deutschland: Messung kognitiver Fähigkeiten und psychologischer Kompetenzen. Bundesgesundheitsblatt, 61, 178–186.

Somaa, F., Asghar, A., & Hamid, P. F. (2021). Academic performance and emotional intelligence with age and gender as moderators: A meta-analysis. Developmental Neuropsychology, 46, 537–554.

Sommer, M., & Arendasy, M. (2015). Further evidence for the deficit account of the test anxiety–test performance relation-ship from a high-stakes admission testing setting. *Intelligence*, 53, 72–80.

Sommer, M., & Arendasy, M. (2016). Does trait test anxiety compromise the measurement fairness of high-stakes scholastic achievement tests? *Learning and Individual Differences*, 50, 1–10.

Sommer, M., Arendasy, M., Punter, J.-F., Feldhammer-Kahr, M., & Rieder, A. (2019). Do individual differences in test-takers' appraisal of admission testing compromise measurement *fairness? Intelligence*, 73, 16–29.

Spicher B. (2018). EMS – Eignungstest für das Medizinstudium in der Schweiz 2018: Bericht 25 über die Durchführung und Ergebnisse 2018. Fribourg: University of Fribourg.

Spicher B. (2019). EMS – Eignungstest für das Medizinstudium in der Schweiz 2019: Bericht 26 über die Durchführung und Ergebnisse 2019. Fribourg: University of Fribourg.

Spicher B. (2020). EMS – Eignungstest für das Medizinstudium in der Schweiz 2020: Bericht 28 über die Durchführung und Ergebnisse 2020. Fribourg: University of Fribourg.

Spicher B. (2021). EMS – Eignungstest für das Medizinstudium in der Schweiz 2021: Bericht 29 über die Durchführung und Ergebnisse 2021. Fribourg: University of Fribourg.

Spiel, C., Schober, B., & Litzenberger, M. (2008). Evaluation der Eignungstests für das Medizinstudium in Österreich. Zusammenfassung und Empfehlungen. Universität Wien: Fakultät für Psychologie.

Stegers-Jager, K. M. (2017). Lessons learned from 15 years of non-grades-based selection for medical school. Medical Education, 52, 86–95.

Stegt, S. J., Didi, H.-J., Zimmerhofer, A., & Seegers, P. K. (2018). Akzeptanz von Auswahlverfahren zur Studienplatzvergabe. Zeitschrift für Hochschulentwicklung, 13, 15–35.

van der Linden, W. J. & Hambelton, R. K. (1997). Handbook of modern Item Response Theory. Springer.

Testkuratorium (2006). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deut-scher Psychologenvereinigungen. Report Psychologie, 31, 492–499.

Testkuratorium (2007). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deut-scher Psychologenvereinigungen. Psychologische Rundschau, 58, 25–30.

Testkuratorium (2018). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deut-scher Psychologenvereinigungen. Revidierte Fassung. Psychologische Rundschau, 58, 25–30.

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs-eine Metaanalyse. Zeitschrift für Pädagogische Psychologie, 21, 11–27.

Trost, G. (1989). A nationwide testing program for admission to medical schools in West Germany. In: King, R. C. & Collins, J. K. (Eds.), Social applications and issues in psychology (131–137). Amsterdam: Elsevier Science Publishers.

Trost, G., Klieme, E. & Nauels, H.-U. (1997). Prognostische Validität des Tests für medizinische Studiengänge (TMS). In T. Herrmann (Hrsg.), Hochschulentwicklung - Aufgaben und Chancen (S. 57–78). Heidelberg: Roland Asanger.

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: Institut für Test- und Begabungsforschung.

Vetter, M. & Sommer, M. (2012). Ergebnisdokumentation zur prognostischen Validität des PMU Aufnahmeverfahrens. Mödling: Schuhfried.

Xi, X. (2010). How do we go about investigating test fairness? Language Testing, 27, 147–170.

Ziegler, M. & Bühner, M. (2012). Grundlagen der Psychologischen Diagnostik. Springer.

Zimmermann, S., Klusmann, D., & Hampe, W. (2018). Angleichung von Schulnoten für die Studierendenauswahl. Zeitschrift für Hochschulentwicklung, 13, 179–197.