Martin ARENDASY[1], Markus SOMMER, Martina FELDHAMMER-KAHR, H. Harald FREUDENTHALER, Fritz Joachim PUNTER & Anita RIEDER (Graz)

Fairness als zentrale Herausforderung moderner Aufnahmeverfahren

Zusammenfassung

In der Praxis wird zumeist angenommen, dass die Ergebnisse in einem Aufnahmeverfahren individuelle Unterschiede in den zu erfassenden Fähigkeiten fair messen. Der vorliegende Beitrag gibt einen Überblick über Modelle zur Fairness von Aufnahmeverfahren und deren Überprüfung in der Praxis anhand von Beispielen zum Aufnahmeverfahren für Medizinische Studiengänge in Österreich (MedAT). Der Beitrag arbeitet zudem auch den engen Zusammenhang zwischen Fairness und anderen Gütekriterien heraus und veranschaulicht die Vorteile einer stark theoriegeleiteten automatischen Itemgenerierung (AIG), um auf die erhöhten Qualitätsanforderungen bei modernen Aufnahmeverfahren reagieren zu können.

Schlüsselwörter

Aufnahmeverfahren, Fairness, automatisierte Itemgenerierung

[1] E-Mail: martin.arendasy@uni-graz.at

Fairness as a central challenge of modern admissions tests

Abstract

In practice, it is often assumed that admissions test scores constitute valid and fair measures of individual differences in the latent traits such tests are designed to measure. The present article discusses facets of fairness relevant to admissions testing and describes how these challenges can be met, using studies on the fairness of the Austrian medical school admission test (MedAT) as an example. In doing so, the article also underscores the relevance of a more theory-based construction of admissions tests using current models of automatic item generation (AIG) to meet the various challenges of modern admissions testing.

Keywords

admissions testing, fairness, automated item generation

1 Einleitung

In den letzten Jahren zeigte sich ein vermehrtes Interesse an Aufnahmeverfahren. Dieses Interesse liegt nicht nur an ihrem metaanalytisch belegten Nutzen zur Vorhersage des späteren Studienerfolgs (z. B. DONNON, PAOLUCCI & VIOLATO, 2007; HELL, TRAPMANN & SCHULER, 2007; SACKETT, KUNCEL, ARNESON, COOPER & WATERS, 2009), sondern auch an ihrer höheren Ökonomie und Akzeptanz im Vergleich zu anderen Auswahlmethoden (HELL & SCHULER, 2005). Da Aufnahmeverfahren Wettbewerbssituationen darstellen, kommt der Fairness des Aufnahmeprozesses eine zentrale Bedeutung zu (vgl. AERA, APA & NCME, 2014; BORSBOOM, ROMEIJN & WICHERTS, 2008; CAMILLI, 2013; KUNNAN, 2000; MILLSAP, 2011; MISLEVY et al., 2013; WAINER, 2002; XI, 2010). Die folgenden Kapitel geben einen Überblick über aktuelle Definitionen von Fairness und versuchen, die Bedeutung dieses Gütekriteriums anhand von Beispielen zum Aufnahmeverfahren für Medizinische Studiengänge in Österreich (MedAT) herauszuarbeiten.

2 Definition von Fairness

Das Konzept der Fairness umfasst verschiedene Facetten, die teilweise aufeinander aufbauen und in enger Beziehung zu anderen Gütekriterien stehen (vgl. XI, 2010; KANE, 2010; KUNNAN, 2000). In Anlehnung an die amerikanischen Standards für pädagogisches und psychologisches Testen (AERA, APA & NCME, 2014) unterscheiden wir folgende Facetten: (1) die Gleichbehandlung aller Bewerber/innen, (2) die Bereitstellung gleicher Möglichkeiten, sich mit den Anforderungen des Aufnahmeverfahrens vertraut zu machen, sowie (3) den Nachweis einer fehlenden systematischen Benachteiligung.

2.1 Fairness als Gleichbehandlung aller Bewerber/innen

Das Konzept der Fairness als Gleichbehandlung aller Bewerber/innen ist eng mit dem Gütekriterium der Durchführungs- und Auswertungsobjektivität verknüpft (vgl. XI, 2010; KANE, 2010; KUNNAN, 2000). Konkret soll sichergestellt werden, dass alle Bewerber/innen unter vergleichbaren Bedingungen getestet werden, und dass die Bewertung ihrer Antworten nach einem eindeutigen Lösungsschlüssel erfolgt. In der Regel wird diese Facette der Fairness durch einen hohen Grad an Standardisierung der Durchführung und Auswertung sichergestellt (KUNNAN, 2000). Das Ziel besteht darin, allen Bewerberinnen/Bewerbern die gleiche Chance zu geben, ihre Kompetenz unter Beweis zu stellen. Hierzu können auch Adaptierungen des Aufnahmeverfahrens für einzelne Personengruppen (z. B. Personen mit speziellen Bedürfnissen) erforderlich sein. Die Äquivalenz der Originalform und der adaptierten Form muss jedoch empirisch überprüft werden, um die Fairness des Aufnahmeverfahrens für unterschiedliche Personengruppen zu gewährleisten.

2.2 Fairness als gleiche Vorinformation für alle Bewerber/innen

Die amerikanischen Standards sehen in den gleichen Möglichkeiten aller Bewerber/innen, sich mit den Inhalten und dem Ablauf des Aufnahmeverfahrens vertraut zu machen, einen weiteren zentralen Aspekt der Fairness (AERA, APA & NCME,

2014). Diese Facette der Fairness kann in der Praxis durch Testwiederholungen und Testcoachings beeinträchtigt werden.

2.2.1 Testwiederholungen als Beeinträchtigung der Fairness

Die Möglichkeit, ein Aufnahmeverfahren zu wiederholen, zählt zur gängigen Praxis (LIEVENS, BUYSE & SACKETT, 2005). Aktuelle Metaanalysen zeigen, dass es durch wiederholte Antritte zu einer Verbesserung der Testscores bei Leistungstests kommt (vgl. HAUSKNECHT, HALPERT, DI PAOLO & MORIARTY GERRARD, 2007; KULIK, KULIK & BANGERT, 1984; SCHARFEN, PETERS & HOLLING, 2018). Wie stark sich die Testleistung verbessert, hängt unter anderem vom zeitlichen Abstand zwischen den Antritten ab. Generell fallen Verbesserungen der Testleistung bei kürzeren Zeitabständen höher aus. Sie sind jedoch auch nach über fünf Jahren noch nachweisbar. Die vorliegenden Metaanalysen zeigen zudem, dass die Zugewinne in der Testleistung bei identischen Testaufgaben höher ausfallen (0.40 bis 0.42 Standardabweichungseinheiten) als bei Parallelformen mit neu konstruierten Testaufgaben (0.22 bis 0.23 Standardabweichungseinheiten). Die berichteten Zugewinne in der Testleistung durch Testwiederholungen scheinen aktuellen Befunden zufolge durch ein Einüben testspezifischer Bearbeitungsstrategien zustande zu kommen (ARENDASY & SOMMER, 2013a; LIEVENS, REEVE & HEGGESTAD, 2007). Aus Sicht der Fairness stellen Testwiederholungen ein Problem dar, da mit der Verbesserung der Testleistung durch ein Einüben testspezifischer Bearbeitungsstrategien ein Vorteil für Personen mit mehrfachen Antritten verbunden ist. Eine Möglichkeit, mit diesem Problem umzugehen, besteht darin, allen Bewerberinnen/Bewerbern die Möglichkeit zu bieten, testspezifische Bearbeitungsstrategien vor dem eigentlichen Antritt zum Aufnahmeverfahren einzuüben. Dies kann realisiert werden, indem man allen Bewerberinnen/Bewerbern Probetests und Infobroschüren mit Beispielaufgaben kostenfrei zur Verfügung zu stellt, wie dies beispielsweise beim Aufnahmeverfahren für Medizinische Studiengänge in Österreich (MedAT) gemacht wird. Ergänzend zu dieser Maßnahme wurden im Rahmen der Entwicklung und Evaluation des Aufnahmeverfahrens für Medizinische Studiengänge in Österreich (MedAT) auch eine Reihe anderer Maßnahmen empirisch evaluiert, mit deren Hilfe das Problem der Verbesserung der Testleistun-

gen durch eine Testwiederholung reduziert werden kann. Beispielsweise untersuchten Arendasy und Sommer (2013a), ob sich Zugewinne in der Testleistung durch die Art der Konstruktion der Paralleltestformen weiter reduzieren lassen. Die Autoren griffen hierbei auf die Methode der automatisierten Itemgenerierung (AIG: IRVINE & KYLLONEN, 2002; ARENDASY & SOMMER, 2011, 2012a) zurück. Im Rahmen dieses Ansatzes wird ausgehend von einer präzisen Definition des zu erfassenden latenten Trait ein sogenanntes Kognitives Modell erarbeitet. Dieses spezifiziert die lösungsrelevanten kognitiven Prozesse bei der Bearbeitung der Testaufgaben, und arbeitet zugleich auch Gestaltungsmerkmalen der Aufgaben heraus, die einen Einfluss auf diese Prozesse haben sollen, und dadurch die Schwierigkeit der Testaufgaben beeinflussen (für weitere Details: ARENDASY & SOMMER, 2011, 2012a). Im Rahmen dieses Konstruktionsansatzes bieten sich im Prinzip zwei Möglichkeiten, Parallelformen für eine wiederholte Testvorgabe zu konstruieren. Beispielsweise können Oberflächenmerkmale der Testaufgaben, die sich nachweislich nicht auf deren Schwierigkeit auswirken, ausgetauscht werden. Dies ist der in der Praxis wohl am häufigsten verwendete Ansatz. Andererseits können jedoch auch schwierigkeitsbestimmende Merkmale der Testaufgaben neu kombiniert werden, sodass vergleichbar schwierige Paralleltestaufgaben entstehen. Arendasy und Sommer (2013a) konnten zeigen, dass durch den letztgenannten Ansatz zur Konstruktion von Paralleltests eine weitere Reduktion der Zugewinne in der Testleistung durch wiederholte Testvorgaben erzielt werden kann. Dieser Ansatz wird aktuell auch im Rahmen des Aufnahmeverfahrens für Medizinische Studiengänge in Österreich (MedAT) verwendet. Eine weitere Möglichkeit, einen Vorteil durch wiederholte Antritte zu minimieren, besteht in der Verwendung Computergestützter Adaptiver Tests (CAT). Da CATs sich an die Leistungsfähigkeit der Bewerber/innen anpassen, werden diese stets optimal gefordert. Dies sollte auch die Möglichkeit der Bewerber/innen reduzieren, während der Testbearbeitung zu lernen. In Übereinstimmung mit dieser Hypothese konnten Arendasy und Sommer (2017) zeigen, dass Zugewinne in der Testleistung vernachlässigbar gering ausfallen, wenn die Bewerber/innen zum ersten Messzeitpunkt einen Computergestützten Adaptiven Test bearbeiteten (<0.10 Standardabweichungseinheiten). Die drei hier genannten Methoden zur Reduktion einer systematischen Bevorzugung

einzelner Bewerber/innengruppen stellen einander ergänzende Maßnahmen dar. Sie stellen jedoch hohe Anforderungen an die Testkonstruktion, die durch klassische Methoden der Testkonstruktion nur schwer realisierbar sind (vgl. HORNKE & HABON, 1986).

2.2.2 Testcoachings als Beeinträchtigung der Fairness

Kommerzielle Anbieter/innen von Testcoachings werben oft mit dem Versprechen, dass sich durch die Teilnahme am Coaching die Chancen, aufgenommen zu werden, deutlich erhöhen. Aktuelle Befunde zeigen jedoch, dass der Vorteil eines kommerziellen Testcoachings gegenüber kostenfrei angebotenen Möglichkeiten zur Vorbereitung eher gering ausfällt (z. B. ALLALOUF & BEN-SHAKHAR, 1998; ARENDASY, SOMMER, GUTIÉRREZ-LOBOS & PUNTER, 2016; BANGERT-DROWNS, KULIK & KULIK, 1983; BECKER, 1990; BRIGGS, 2004, 2009; MESSICK & JUNGEBLUT, 1981; POWERS, 1988, 2012; POWERS & ROCK, 1999; WITT, 1993). Die Frage, wie viel Zeit und Mühe jemand in die Vorbereitung investiert, scheint entscheidender zu sein als die Frage, auf welche Vorbereitungsmöglichkeiten man dabei zurückgreift. Dies zeigte sich auch in zwei unabhängigen Studien zur Evaluation des Einflusses kommerzieller und nichtkommerzieller Testvorbereitungsangebote für das Aufnahmeverfahren für Medizinische Studiengänge in Österreich (ARENDASY et al., 2016; ARENDASY, SOMMER & FELDHAMMER, 2016b). In diesen beiden Studien konnte zudem gezeigt werden, dass es durch Unterschiede in der Nutzung verschiedener Angebote zur Vorbereitung zu keiner systematischen Benachteiligung einzelner Gruppen an Bewerberinnen/Bewerbern kommt. Das Problem kommerzieller Testcoachings verdeutlicht jedoch die Notwendigkeit, nicht kommerzielle Vorbereitungsangebote seitens der Hochschulen kostenfrei zur Verfügung zu stellen, was mit einem Mehraufwand bei der Konstruktion des Aufnahmeverfahrens verbunden ist.

2.3 Fairness als fehlende systematische Benachteiligung

Diese Facette der Fairness steht in einem engen Zusammenhang mit der Konstrukt- und Kriteriumsvalidität (vgl. CAMILLI, 2013; KANE, 2010; KUNNAN, 2000; XI,

2010). Generell unterscheidet man zwischen (1) Messfairness, (2) Prognosefairness und (3) Selektionsfairness. Die hierarchische Beziehung dieser drei Facetten der psychometrischen Fairness ist in Abbildung 1 veranschaulicht.

Abb. 1: Hierarchische Beziehung einzelner Facetten der Fairness

2.3.1 Messfairness bzw. Messinvarianz

Die amerikanischen Standards für pädagogisches und psychologisches Testen fordern, dass empirisch belegt werden muss, dass es durch einzelne Aufgaben zu keiner systematischen Bevorzugung oder Benachteiligung einzelner Personengruppen kommt (AERA, APA & NCME, 2014). In der Fachliteratur spricht man hier von Messfairness. Messfairness bedeutet, dass Personen mit gleicher Fähigkeit, unabhängig von ihrer Gruppenzugehörigkeit (z. B. Geschlecht, sozioökonomischer Status etc.), die gleiche Wahrscheinlichkeit haben, eine Aufgabe zu lösen (vgl. BORSBOOM et al., 2008; CAMILLI, 2013; KUNNAN, 2000; MILLSAP, 2011; MISLEVY et al., 2013; WAINER, 2002; XI, 2010). Dies bedeutet, dass die Schwierigkeitsrelationen der Aufgaben für alle Personen gleich bzw. invariant sein müssen. Diese Annahme kann mit Methoden der Item Response Theorie (IRT: DE BOECK & WILSON, 2004; MILLSAP, 2011; ROST, 2004) oder der Konfirmatorischen Faktorenanalyse (CFA: MILLSAP, 2011) empirisch geprüft werden. Die Gemeinsamkeiten und Unterschiede zwischen diesen beiden Ansätzen wurden von verschiedenen Autorinnen/Autoren herausgearbeitet (zusammenfassend: McDONALD, 1999). Aktuell geht man davon aus, dass sich diese Ansätze ideal ergänzen, da sie jeweils für unterschiedliche Störeinflüsse der Messfairness sensitiv zu sein scheinen (für praktische Beispiele: ARENDASY & SOMMER, 2012b, 2013c).

Sprechen die empirischen Befunde gegen die Annahme der Messfairness, kommt es zu einer systematischen Bevorzugung oder Benachteiligung einzelner Gruppen an Bewerberinnen/Bewerbern. Dies wirkt sich auch nachteilig auf die Validität und Fairness der Rangreihung der Bewerber/innen aus, da mehr Personen aus der bevorzugten Gruppe aufgenommen werden, als es aufgrund ihrer wahren Fähigkeit gerechtfertigt wäre (vgl. BORSBOOM et al., 2008; MILLSAP & KWOK, 2004).

Im Gegensatz zu den USA sind Studien zur Überprüfung der Messfairness im deutschsprachigen Raum noch eher selten. In der Praxis wird häufig angenommen, dass Messfairness gegeben ist. Empirische Befunde zeigen jedoch, dass diese Annahme auch bei etablierten Aufnahmeverfahren nicht immer zutrifft (z. B. SPIEL, SCHOBER & LITZENBERGER, 2008). Aus diesem Grund kommt der empirischen Überprüfung der Messfairness ein zentraler Stellenwert zu. Im Rahmen der Entwicklung des Aufnahmeverfahrens für Medizinische Studiengänge in Österreich (MedAT) wurde die Fairness einzelner Aufgabengruppen bereits in Vorstudien vor der eigentlichen Entwicklung des Aufnahmeverfahrens untersucht. In einer Reihe von Studien wurden Gestaltungsmerkmale von Aufgaben und Aufgabenformaten hinsichtlich ihrer Wahrscheinlichkeit evaluiert, einzelne Gruppen an Bewerberinnen/Bewerbern systematisch zu benachteiligen (z. B. ARENDASY & SOMMER, 2010, 2011, 2012a, 2012b, 2013b, 2013c; ARENDASY, SOMMER & GITTLER, 2010; ARENDASY, SOMMER, HERGOVICH & FELDHAMMER, 2011; ARENDASY, SOMMER & MAYR, 2012). Die aus diesen Studien gewonnenen Erkenntnisse lieferten wichtige Hinweise für die Gestaltung des Aufnahmeverfahrens. Entsprechend zeigten die jährlich durchgeführten Überprüfungen der Messfairness, dass die Aufgaben des Aufnahmeverfahrens hinsichtlich einer Reihe von Personenmerkmalen als fair angesehen werden können. Zu diesen Personenmerkmalen zählten das Geschlecht, die Staatsbürgerschaft der Bewerber/innen, der Bildungsstand der Eltern, der zuvor besuchte Schultyp der Bewerber/innen, ihr sozioökonomischer Status, die Anzahl an Testwiederholungen, die Art der Vorbereitung auf das Aufnahmeverfahren, die Muttersprache der Bewerber/innen, das Ausmaß an erlebter Testangst während des Aufnahmeverfahrens und der Standort

der Testdurchführung (vgl. ARENDASY, SOMMER & FELDHAMMER, 2013, 2014, 2015, 2016a, 2016b, 2017; SOMMER & ARENDASY, 2015, 2016).

Da eine Verletzung der Messfairness zu zahlreichen Problemen führt, wird häufig empfohlen, messunfaire Aufgaben auszuschließen oder gruppenspezifische Aufgabenschwierigkeiten zu verwenden, um die Fähigkeit der Bewerber/innen zu schätzen (zusammenfassend: DE BOECK & WILSON, 2004; ROST, 2004; MILLSAP, 2011). Dies steht jedoch im Widerspruch zur Fairness im Sinne der Gleichbehandlung aller Bewerber/innen. Ignoriert man hingegen eine Verletzung der Messfairness, kommt es zu einer systematischen Bevorzugung einzelner Gruppen an Bewerberinnen/Bewerbern. Dies verdeutlicht die Notwendigkeit empirischer Studien zur Messfairness und der Berücksichtigung von Überlegungen zur Messfairness im Rahmen der Konzeption des Aufnahmeverfahrens.

2.3.2 Prognosefairness und Selektionsfairness

Unter Prognosefairness versteht man die Gleichheit der Parameter eines Vorhersagemodells für alle Gruppen an Bewerberinnen/Bewerbern (AGUINIS, CULPEPPER & PIERCE, 2010; BORSBOOM et al., 2008; CLEARY, 1968; MEADE & TONIDANDELL, 2010). Die Prognosefairness ist daher eng mit dem Gütekriterium der prognostischen Validität verknüpft (KANE, 2010). Die empirische Überprüfung der Prognosefairness erfolgt zumeist mit Hilfe einer hierarchischen Regression. In einem ersten Schritt werden alle als relevant erachteten Prädiktoren in das Vorhersagemodell aufgenommen. In zweiten Schritt wird zusätzlich die Gruppenzugehörigkeit der Personen in das Modell miteinbezogen. Wenn ein uniformer Prognosebias vorliegt, führt der zweiten Schritt zu einer signifikant besseren Passung des Vorhersagemodells und das Gewicht für die Gruppenzugehörigkeit wird signifikant. Im dritten und letzten Schritt werden schließlich auch die Wechselwirkung(en) zwischen Prädiktor(en) und der Gruppenzugehörigkeit miteinbezogen. Dieser letzte Schritt prüft das Vorhandensein eines nicht-uniformen Prognosebias. Nach Meade und Tonidandel (2010) kommt es in der Praxis häufig nur zu einem uniformen Prognosebias, was teilweise mit der geringeren Sensitivität der Modelltests für einen nicht-uniformen Prognosebias erklärt werden kann (vgl. AGUINIS

et al., 2010). Die Interpretation von Befunden zur Prognosefairness gestaltet sich jedoch häufig schwierig. Empirische Befunde und Simulationsstudien zeigen, dass Ergebnisse zur Prognosefairness nur dann eindeutig interpretierbar sind, wenn sowohl für die Prädiktoren als auch für das Kriterium Messfairness angenommen werden kann (vgl. BORSBOOM et al., 2008; MILLSAP & KWOK, 2004), und das Aufnahmeverfahren zugleich alle relevanten Prädiktoren abdeckt (z. B. FISCHER, SCHULT & HELL, 2013; MATTERN, SANCHEZ & NDUM, 2017; MEADE & FETZER, 2009; SACKETT, LACZO & LIPPE, 2003). Trifft dies nicht zu, besteht die Gefahr, dass die Ergebnisse einen Prognosebias nahelegen, obwohl keiner vorhanden ist, oder einen tatsächlich vorhandenen Prognosebias verschleiern. Dies bedeutet, dass Prognosefairness immer nur im Hinblick auf ein bestimmtes Kriterium und eine bestimmte Auswahl an Prädiktoren beurteilt werden kann.

Die Selektionsfairness baut auf der Annahme der Mess- und Prognosefairness auf. Unter Selektionsfairness versteht man die Gleichheit der Treffsicherheit der Prognosen für alle Gruppen an Bewerberinnen/Bewerbern (BORSBOOM et al., 2008). Hierfür ist es erforderlich, den Anteil der zu Recht und zu Unrecht ausgewählten bzw. abgelehnten Personen anhand der Taylor-Russel-Tafel zu bestimmen. Aktuelle Simulationsstudien zeigen, dass Selektionsfairness nahezu unmöglich zu erreichen ist, wenn trotz Messfairness für die Prädiktoren und die Kriterien ein uniformer Prognosebias vorliegt (BORSBOOM et al., 2008; MILLSAP & KWOK, 2004). In einem solchen Fall sollte daher immer geprüft werden, ob die mangelnde Prognosefairness durch fehlende Prädiktoren im Vorhersagemodell erklärbar ist. Empirische Arbeiten zur Prognose- und Selektionsfairness, deren Ergebnisse eindeutig interpretierbar sind, sind weltweit noch eher selten. Die bisher vorliegenden Befunde unterstreichen jedoch die Bedeutung der Messfairness und einer sorgfältigen Konstruktion der Aufgaben eines Aufnahmeverfahrens für die Erreichung des Ziels der Prognose- und Selektionsfairness. Wie bereits deutlich wurde, bietet die automatisierte Itemgenerierung (AIG) einen attraktiven Ansatz zur Konstruktion von Aufnahmeverfahren, mit dem Anforderungen an die Fairness und Validität von Aufnahmeverfahren bereits während des Prozesses der Testkonstruktion berück-

sichtigt werden können (zusammenfassend: ARENDASY & SOMMER, 2011, 2012a; IRVINE & KYLLONEN, 2002).

3 Fazit

Da Aufnahmeverfahren eine Wettbewerbssituation darstellen, kommt ihrer Fairness eine zentrale Bedeutung zu (vgl. AERA, APA & NCME, 2014). Der Begriff Fairness bezeichnet unterschiedliche Facetten, die eng mit anderen Gütekriterien assoziiert sind und zum Teil hierarchisch aufeinander aufbauen. Generell geht es darum sicherzustellen, dass alle Bewerber/innen die gleichen Möglichkeiten haben, ihre Eignung für ein Studium unter Beweis zu stellen, und dass es weder durch die Gestaltung der Testaufgaben noch durch Unterschiede in den Vorabinformationen der Bewerber/innen zu einer systematischen Bevorzugung oder Benachteiligung einzelner Gruppen an Bewerberinnen/Bewerbern kommt. Die Überprüfung und Sicherstellung der Fairness eines Aufnahmeverfahrens geht dabei weit über den eigentlichen Prozess der Testkonstruktion hinaus und stellt Anbieter/innen und Entwickler/innen von Aufnahmeverfahren vor neuen Herausforderungen, die mit klassischen Methoden der Testkonstruktion nur mehr schwer zu bewältigen sind (HORNKE & HABON, 1986). Eine attraktive Lösung für dieses Problem stellen unterschiedliche Methoden der automatisierten Itemgenerierung (AIG) dar, mit deren Hilfe nicht nur eine hinreichend hohe Anzahl qualitativ hochwertiger Testaufgaben konstruiert werden kann, sondern auch kostenfrei verfügbare Übungsaufgaben, mit deren Hilfe vorab bestehende Unterschiede in den Vorabinformationen der Bewerber/innen reduziert werden können (für einen Überblick zur AIG: ARENDASY & SOMMER, 2011, 2012a; IRVINE & KYLLONEN, 2002).

4 Literaturverzeichnis

Aguinis, H., Culpepper, S. A. & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*, 648-680.

Allalouf, A. & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*, 31-47.

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. (2010). Evaluating the contribution of different item design features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence, 38*, 574-581.

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). Göttingen: Hogrefe.

Arendasy, M. & Sommer, M. (2012a). 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. (2012b). Gender differences in figural matrices: The moderating role of item design features. *Intelligence*, *40*, 584-597.

Arendasy, M. & Sommer, M. (2013a). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41*, 181-192.

Arendasy, M. & Sommer, M. (2013b). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence, 41*, 234-243.

Arendasy, M. & Sommer, M. (2013c). *Automatic item generation and first evidences on the dimensionality. Measurement fairness and construct representation of a picture completion task*. Unpublished research report. Graz: University of Graz.

Arendasy, M. & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence*, *62*, 89-98.

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

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

Arendasy, M., Sommer, M. & Feldhammer, M. (2015). *MedAT-H & MedAT-Z 2015: Auswertungsdokumentation und psychometrische Evaluatio*n. Graz: Universität Graz.

Arendasy, M., Sommer, M. & Feldhammer, M. (2016a). *MedAT-H & MedAT-Z 2016: Auswertungsdokumentation und psychometrische Evaluatio*n. Graz: Universität Graz.

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

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

Arendasy, M., Sommer, M. & Gittler, G. (2010). Combining automatic item generation and experimental designs to investigate the contribution of cognitive components to the gender difference in mental rotation. *Intelligence, 38*, 506-512

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., Hergovich, A. & Feldhammer, M. (2011). Evaluating the impact of depth cue salience in working three-dimensional mental rotation tasks by means of psychometric experiments. *Learning and Individual Differences, 21*, 403-408.

Arendasy, M., Sommer, M. & Mayr, F. (2012). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Cross Cultural Psychology, 43*, 464-479.

Bangert-Drowns, R. L., Kulik, J. A. & Kulik, Ch.-L. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, *53*, 571-585.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373-417.

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

Briggs, D. C. (2004). Evaluating SAT coaching: Gains, effects, and self-selection. In R. Zwick (Hrsg.), *Rethinking the SAT: The future of standardized testing in university admissions* (S. 217-233). New York: Routledge Falmer.

Briggs, D. C. (2009). *Preparation for college admission exams. (NACAC discussion paper)*. Arlington, VA: National Association for College Admission Counseling.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*, 104-120.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

De Boeck, P. & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

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.

Fischer, F. T., Schult, J. & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*, 478-488.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T. & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Hell, B. & Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. *Empirische Pädagogik, 19*, 361-376.

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

Irvine, S. H. & Kyllonen, P. C. (2002). *Item Generation for Test Development*. New Jersey: Lawrence Erlbaum Associates.

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

Kulik, J. A., Kulik, Ch.-L. C. & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435-447.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Hrsg.), *Fairness and validation in language assessment* (S. 1-14). Cambridge, UK: Cambridge University Press.

Lievens, F., Buyse, T. & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981-1007.

Lievens, F., Reeve, C. L. & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*, 1672-1682.

Mattern, K., Sanchez, E. & Ndum, E. (2017). Why do achievement measures underpredict female academic performance? *Educational Measurement: Issues and Practice*, *36*, 47-57.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meade, A. W. & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods, 12*, 738-761.

Meade, A. W. & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology, 3*, 192-205.

Messick, S. & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191-216.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: 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.

Powers, D. E. (1988). *Preparing for the SAT: A survey of programs and resources* (ETS RR-88-40). New York: College Board.

Powers, D. E. (2012). *Understanding the impact of special preparation for admissions tests*. ETS research report series.

Powers, D. E. & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning scores. *Journal of Educational Measurement*, *36*, 93-118.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion*. Bern: Huber.

Sackett, P. R., Kuncel, N. R., Areson, J. J., Cooper, S. R. & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions test and post-secondary academic performance? *Psychological Bulletin, 135*, 1-22.

Sackett, P. R., Laczo, R. M. & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88*, 1046-1056.

Scharfen, J., Peters, J. M. & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44-66.

Sommer, M. & Arendasy, M. (2015). Further evidence for the deficit account of the test anxiety-test performance relationship 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.

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

Wainer, H. (2002). On the automatic generation of items: Some whens, whys and hows. In S. H. Irvine & P. C. Kyllonen (Hrsg.), *Item generation for test development* (S. 287-316). New Jersey: Lawrence Erlbaum.

Witt, E. A. (1993). *Meta-analysis and the effects of coaching for aptitude tests*. Paper presented at the Annual meeting of the American Educational Research Association.

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

Autorinnen/Autoren

Univ.-Prof. Dr. Martin ARENDASY || Universität Graz, Institut für Psychologie || Universitätsplatz 2, A-8010 Graz

https://psychologie.uni-graz.at/de/psychologische-diagnostik-und-methodik/team/

martin.arendasy@uni-graz.at

Dr. Markus SOMMER || Universität Graz, Institut für Psychologie || Universitätsplatz 2, A-8010 Graz

https://psychologie.uni-graz.at/de/psychologische-diagnostik-und-methodik/team/

markus.sommer@uni-graz.at

Dr. Martina FELDHAMMER-KAHR || Universität Graz, Institut für Psychologie || Universitätsplatz 2, A-8010 Graz

https://psychologie.uni-graz.at/de/psychologische-diagnostik-und-methodik/team/

martina.feldhammer@uni-graz.at

Ao.-Univ.-Prof. Dr. H. Harald FREUDENTHALER || Universität Graz, Institut für Psychologie || Universitätsplatz 2, A-8010 Graz

https://psychologie.uni-graz.at/de/psychologische-diagnostik-und-methodik/team/

heribert.freudenthaler@uni-graz.at

Mag. Joachim Fritz PUNTER || Medizinische Universität Wien, Assessment & Skills || Spitalgasse 23, A-1090 Wien

https://teachingcenter.meduniwien.ac.at/abteilungen/assessment-skills/mitarbeiterinnen/

joachim.punter@meduniwien.ac.at

VR Univ.-Prof. Dr. Anita RIEDER || Medizinische Universität Wien, Zentrum für Public Health || Kinderspitalgasse 15, A-1090 Wien

https://www.meduniwien.ac.at/hp/sozialmedizin/allgemeine-informationen/mitarbeiterinnen/

anita.rieder@meduniwien.ac.at