Intelligenzdiagnostik im Wandel der Technik

Analysen zur Anpassung von Intelligenzdiagnostik für computerisierte Testungen und Studierendenauswahl

Dissertation

Zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

der Fakultät HW

Bereich Empirische Humanwissenschaften

der Universität des Saarlandes

vorgelegt von

Marco Koch

Aus Dudweiler

Saarbrücken, 2023

Dekan:	Prof. Dr. Peter Loos
1. Berichterstatter:	PD Dr. Nicolas Becker
2. Berichterstatterin:	Prof. Dr. Gisa Aschersleben
3. Berichterstatter:	Prof. Dr. Florian Schmitz
Tag der Disputation:	06.02.2023

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AR	Augemented Reality
BCI	Brain-Computer Interface
CDS	Cognitive Design System
DIF	Differentielle Itemfunktion
g	Generalfaktor der Intelligenz
IRT	Item Response Theory
MGCFA	Konfirmatorische Multigruppen-Faktorenanalyse
SES	Sozioökonomischer Status
VR	Virtual Reality

Überblick über die relevanten Studien

Diese Dissertation umfasst insgesamt drei Studien. Die Studien I und II wurden in international anerkannten wissenschaftlichen Zeitschriften (mit Peer-Review) veröffentlich, Studie III wurde zur Veröffentlichung eingereicht und befindet sich derzeit im Begutachtungsprozess.

Die vollständigen Manuskripte zu den Studien I und II liegen in ihrer publizierten Form im Anhang dieser Abhandlung vor. Studie III befindet sich im Anhang in der Form zum Zeitpunkt der Einreichung.

Studie I	Koch, M., Becker, N., Spinath, F. M., & Greiff, S. (2021). Assessing intelligence without intelligence tests. Future perspectives. *Intelligence*, 89. https://doi.org/10.1016/j.intell.2021.101596
Studie II	Levacher, J.*, Koch, M.*, Hissbach, J., Spinath, F. M., & Becker, N. (2021). You can play the game without knowing the rules – But you're better off knowing them: The influence of rule knowledge on Figural Matrices Tests. *European Journal of Psychological Assessment*, https://doi.org/10.1027/1015-5759/a000637 (*joint first authorship)
Studie III	Koch, M., Becker, N., Greiff, S., & Spinath, F. M. (2021). *Development and Validation of the Open Matrices Item Bank.* Manuscript submitted for publication.

1 Einleitung

Die vorliegende Dissertation beschäftigt sich mit der Frage, wie die Intelligenzdiagnostik angepasst werden kann, um neue technische Möglichkeiten des Computerzeitalters im Sinne der Forschung auszuschöpfen und gleichzeitig auf neue gesellschaftliche Situationen und Herausforderungen einzugehen. Zu diesem Zweck wird zunächst in Kapitel 2 eine Übersicht über das Konstrukt Intelligenz und seine Diagnostik gegeben. Daran anschließend wird eine Studie präsentiert, in der aktuelle Entwicklungen der Intelligenzdiagnostik gesichtet und auf deren Basis zukünftige Forschungsschwerpunkte und Veränderungen der Intelligenzdiagnostik behandelt werden (Koch, Becker, et al., 2021). Kapitel 3 stellt einen solche Forschungsschwerpunkt – die computerbasierte Diagnostik – dar und demonstriert anhand einer Studie die Entwicklung und Evaluation einer nicht-kommerzielle Item Bank mit über 200 figuralen Matrizen (Koch, Spinath, et al., 2021). In Kapitel 4 wird die Nutzung von Intelligenztests für Studierendenauswahlverfahren und die damit gestiegenen Anforderungen an Testfairness thematisiert. Die hierbei näher betrachtete Studie zeigt, wie sehr sich bereits minimale Interventionen auf Testergebnisse auswirken (Levacher et al., 2021). Schließlich werden in Kapitel 5 die Zusammenhänge der Studien beleuchtet und die Implikationen für das Forschungsfeld der Intelligenz(-diagnostik) diskutiert.

2 Intelligenz

2.1 Allgemeine Einordnung

Es steht außer Frage, dass es sich bei Intelligenz um eines der zentralsten Konstrukte psychologischer Forschung handelt. So weist Intelligenz nicht nur Zusammenhänge zu Bildungserfolg (Roth et al., 2015) und Berufserfolg (Schmidt & Hunter, 2016, 1998) auf sondern erstreckt sich über alle Bereich des täglichen Lebens (Gottfredson, 1997, 2004; Neisser et al., 1996). Aktuelle Befunde zur Struktur der Intelligenz und beteiligter Prozesse basieren auf einer langen Forschungstradition, die sich von den frühen theoretischen Überlegungen Spearmans bis hin zu heutigen sehr differenzierten Intelligenzmodellen (McGrew, 2009) erstreckt. Über diese Zeitspanne hinweg ist auch eine Vielzahl an Definitionen der Intelligenz entstanden. Heutzutage herrscht jedoch Konsens, dass es sich um eine breite kognitive Fähigkeit handelt, die das Verständnis von komplexen Ideen, die Anpassungsfähigkeit an Umweltbedingungen, das Lernen aus Erfahrung und das Lösen von Problem durch Analyse umfasst (Neisser et al., 1996).

Heutzutage wird in der Regel angenommen, dass der Intelligenz eine hierarchische Struktur zugrunde liegt (vgl. Abb. 1 ; McGrew, 2009). Hierbei steht an oberster Stelle der Generalfaktor der Intelligenz (*g*), der bereits von Spearman (Spearman, 1904) als gemeinsamer Varianzanteil mehrerer psychologischer Leistungstests identifiziert wurde. Auf der nächsten Hierarchieebene finden sich breite Intelligenzfaktoren wie die kristalline und fluide Intelligenz aus Cattells Forschung (Cattell, 1963) sowie weitere Fähigkeiten der Wahrnehmung, Informationsverarbeitung und -speicherung als auch spezifische sprachliche und numerische Fähigkeiten. Darunter wiederum finden sich auf einer dritten Ebene sehr eng umschrieben kognitive Fähigkeiten wie etwa Schreibgeschwindigkeit, Gedächtniskapazität und Wahrnehmungsgeschwindigkeit (McGrew, 2009).

Aufgrund dieser langen Forschungstradition und der praktischen Bedeutsamkeit der Intelligenz wurden auch eine Vielzahl an Intelligenztestverfahren entwickelt. Jedoch basieren diese auch heutzutage meist auf der Diagnostik mit Papiertestungen. Deren Layout und Darstellung dient zeitgleich als Grundlage für die meisten existierenden

computerbasierten Verfahren. Lediglich das Ankreuzen von Antworten wurde modifiziert und erfolgt in diesen computerisierten Tests durch Anklicken des richtigen Kästchens.

Daher sollte der aktuelle Zustand der Intelligenzdiagnostik kritisch betrachtet werden. Dies erfolgt durch Studie I (Koch, Becker, et al., 2021), die mit Hilfe einer Literaturrecherche und Analogieschlüssen aus anderen Forschungsbereichen demonstriert, welche Entwicklungsmöglichkeiten für die Intelligenzdiagnostik in Zukunft vorhanden sind und wie eventuell sogar eine Erfassung der Intelligenz ohne den Einsatz klassischer Testverfahren ermöglicht werden könnte.

2.2 Modernisierung der Intelligenzdiagnostik

Hinführung

Im Laufe des 20. und insbesondere des 21. Jahrhunderts hat der technologische Fortschritt zu großen Veränderungen und weitreichenden Erkenntnisse in der gesamten Wissenschaft geführt. Die stetige Verdopplung der Rechenleistung moderner Computer hat dazu beigetragen, dass heutzutage ein Großteil der Menschheit mit ihrem Smartphone einen Computer bei sich trägt, der mehrere tausendmal leistungsfähiger ist als jener, der zur Steuerung der Apollo-Mission genutzt wurde.

Trotz dieser enormen digitalen Weiterentwicklung und den damit verbundenen Möglichkeiten gab es überraschend wenig Veränderungen im Bereich der Intelligenzdiagnostik. Natürlich ist diese Stagnation zu einem sehr großen Anteil darauf zurückführbar, dass die etablierten Verfahren zur Messung der menschlichen Intelligenz aus psychometrischer Sicht über sehr gute Eigenschaften verfügen und sich vielfach bewährt haben. Dies darf jedoch nicht als Grund verstanden werden, keine Forschung mehr durchzuführen, die zur Verbesserung der Diagnostik beitragen kann.

Ziel dieser Studie

Ziel der Studie I (Koch, Becker, et al., 2021) war es anhand bereits existierender Forschung und Technologien, neue Forschungsansätze zur Verbesserung der Intelligenzdiagnostik zu finden und hinsichtlich Ihrer gegenwärtigen Umsetzbarkeit zu betrachten.

Methodik

Da es sich bei Studie I (Koch, Becker, et al., 2021) um eine theoretische Arbeit handelt stand eine Literaturrecherche im Zentrum der Methodik. Hierzu wurden in einem ersten Schritt aktuelle Intelligenztheorien sowie Intelligenztestverfahren gesichtet. Es wurde anschließend nach modernen Diagnostikansätzen in den Bereichen der Intelligenz- und Persönlichkeitsforschung sowie der Humanmedizin recherchiert, um einen Überblick über bereits umgesetzte Neuerungen zu erhalten. Schließlich wurde auch aktuelle Literatur zu künstlicher Intelligenz und Machine-Learning-Modellen gesammelt, um fundierte Vorhersagen für zukünftige Möglichkeiten treffen zu können.

Aus diesen Datenquellen ergaben sich drei Ansatzpunkte zur Modernisierung der Intelligenzdiagnostik. (1) Der Einsatz von Computern in der psychologischen Diagnostik ist immer noch sehr stark auf eine Vereinfachung der Testadministration beschränkt, anstatt einzigartige Möglichkeiten wie die Erfassung von Prozessdaten auszunutzen. (2) Fortschritte in der künstlichen Intelligenz haben dazu geführt, dass aus recht diffusen Ausgangsdaten sehr präzise Vorhersagen getroffen werden können, die traditionelle Testverfahren überflüssig machen könnten. (3) Forschung aus der Medizin und Neurowissenschaft ermöglichen teilweise die Entwicklung von *brain-computer-interfaces* (BCI; Gehirn-Computer-Schnittstelle), die eine direkte Interaktion von Gehirn und Computer ermöglichen (Allison et al., 2007) und eventuell zu einer direkten Messung der Intelligenz führen könnten.

Ergebnisse

Computerbasierte Diagnostik stellt einen der frühesten Berührungspunkte zwischen Technologie und Intelligenzforschung dar (z.B. English et al., 1977; Hunt & Pellegrino, 1985). Frühe Forschung befasste sich häufig mit der Frage der Äquivalenz von computerbasierten und papierbasierten Testverfahren. In einer Meta-Analyse wurde dann zusammengefasst, dass insgesamt eine beträchtliche Äquivalenz zwischen den Modalitäten besteht ($\rho = .90$), die jedoch bei Tests mit einer starken Speed-Komponente eingeschränkt ist ($\rho = .72$, Mead & Drasgow, 1993; für ähnliche Befunde s. Wang et al., 2007, 2008). Trotz dieser positiven Befunde und einer bereits sehr früh formulierten Forderung Computer nicht auf die Vereinfachung von

Testadministration zu beschränken (Hunt & Pellegrino, 1985) wird weiterhin viel Forschung betrieben, um die Äquivalenz zwischen den Darbietungsmodalitäten zu untersuchen.

Eine methodische Grundlage, um dieser Forderung gerecht zu werden, findet sich im *cognitive design system* (CDS) Ansatz (Abad et al., 2017; Embretson, 1998). Mit diesem Ansatz wird versucht die Konstruktvalidität eines Tests zu beeinflussen, in dem gezielt bestimmte Aspekt der Items variiert werden, bei denen man einen bestimmten kognitiven Zusammenhang erwartet. Primi (2014) folgte diesem Ansatz bei der Entwicklung eines Matrizentests. Dabei nahm er an, dass fluide Intelligenz sich aus den beiden Komponenten Arbeitsgedächtnis und zentrale Exekutive (*goal management* und Abstraktion) zusammensetzt. Dabei können diese beiden Bereiche anhand der Itemdesignvariablen (Quantität von Information, Regelart und perzeptuelle Komplexität) manipuliert werden. Durch die Itemdesignvariablen konnte er in seiner Studie 87% der Varianz in den Itemschwierigkeitsparametern vorhersagen. Durch solche Ansätze wird es möglich, genauere Schlüsse zu ziehen. Zum Beispiel, ob ein schlechter Testwert in einem Intelligenztest durch eine Begrenzung der Arbeitsgedächtniskapazität, eine mangelnde Inhibitionsfähigkeit oder ähnliche Faktoren entstanden ist.

Über bestimmte Entwicklungsansätze hinaus liegt auch großes Potential darin, neue Datenquellen auszuschöpfen. Während Prozessvariablen wie etwa die Reaktionszeit in der Experimentalpsychologie schon seit langem eine zentrale Rolle haben, so werden sie zu diagnostischen Zwecken eher selten genutzt. Dies liegt unter anderem daran, dass sie bei papierbasierten Testungen, fast unmöglich zu erheben sind. In computerbasierten Tests hingegen stellt die Messung von Reaktionszeiten kein Problem dar. Daher wurde bereits mehrfach demonstriert, dass Prozessvariablen auch hohe prädiktive Validität aufweisen (Dodonov & Dodonova, 2012; Greiff et al., 2015; Kroner et al., 2005) und Aufschlüsse über die Grundlagen intelligenten Verhaltens liefern können (Krieger et al., 2019).

Ein weiterer Ansatz wie die Möglichkeiten computerbasierter Diagnostik besser ausgenutzt werden können findet sich im Einsatz von Videospielen (für eine breite Übersicht erfolgreicher Implementationen s. Simons et al., 2021). Ursprünglich wurde dieser Ansatz erforscht, um die Motivation von sich bewerbenden Personen zu steigern.

Dabei wurden hauptsächliche spieleartige Elemente wie Bestenlisten und Fortschrittsbalken genutzt (Hamari et al., 2014). Es wurden auch eine Reihe von *serious games* entwickelt, deren primärer Zweck nicht hedonistischer Natur ist, sondern der Diagnostik und Intervention dienen. Diese haben sich vor allem im HR-Bereich bewährt (Bellotti et al., 2013). Schließlich ist es auch gelungen Intelligenztestwerte mit sehr hoher Genauigkeit aus kommerziellen Videospielen vorherzusagen (Kokkinakis et al., 2017; Quiroga et al., 2015, 2017, 2019). Durch eine Kombination mit der Analyse von Prozessdaten kann die Spielzeit teilweise auf die Hälfte reduziert werden, ohne die prädiktive Validität zu mindern (Guo et al., 2020). Ein nächste Entwicklungsschritt könnte in dieser Forschungsrichtung in der Aufnahme von *augmented realtiy* (AR) oder *virtual reality* (VR) Szenarien bestehen, um naturalistischere Umgebungen zu schaffen und damit eventuell eine höhere externe Validität zu ermöglichen. Zurzeit gibt es jedoch kaum Forschung zur Auswirkung auf das menschliche Gehirn. Bisherige Studien zeigen, dass besonders Kinder nur schlecht zwischen realen und VR Szenarien unterscheiden können (Cadet & Chainay, 2021; Segovia & Bailenson, 2009). Außerdem gibt es Befunde, die für ein gewisses Ausmaß an Neuroplastizität durch VR Szenarien sprechen (Deutsch & Westcott McCoy, 2017).

Als zweiter Ansatz für eine Modernisierung der Intelligenzdiagnostik wurde in Studie I (Koch, Becker, et al., 2021) die künstliche Intelligenz identifiziert. Machine-Learning-Modell sind mittlerweile so weit fortgeschritten, dass sie Aufgaben erfüllen können, die zuvor zwangsläufig von einem Experten abhängig waren. So sind zum Beispiel neuronale Netzwerke mittlerweile genauso valide in der Krebsdiagnose per Röntgenaufnahme wie Ärzte (Rajpurkar et al., 2017). Aber auch bezogen auf die psychologische Diagnostik liegen bereits erste Ergebnisse zu Machine-Learning-Modellen vor. In der Persönlichkeitspsychologie konnte bereits gezeigt werden, dass anhand von Social Media Profilen die Persönlichkeit, abgebildet durch das Big-Five Modell, sehr gut vorhergesagt werden kann (Azucar et al., 2018; Schwartz et al., 2013; Tandera et al., 2017). Im Bereich der Intelligenzforschung ist dieser Ansatz noch weniger erforscht. Eine Meta-Analyse berichtet anhand von $k = 3$ Studien einen relativ geringen Zusammenhang von $r = .29$ (CI [.19 - .36], Settanni et al., 2018). Die inkludierten Studien verfolgten zudem sehr unterschiedlichen Ansätzen (z.B. Analyse von Aktivitäten

vs. Profilbilder als Prädiktoren), die jeweils auf eine spezifisch umschrieben Datenquellen zurückgriffen. Es kann daher angenommen werden, dass eine größeren Inputdatenmenge eine Verbesserung der Prädiktion bedingt.

Bei diesem Ansatz steht außer Frage, dass vor einer Anwendung in der Praxis noch viele methodische und ethische Aspekte betrachtet werden müssen. Die benötigten Ausgangsdaten für ein verlässliches Vorhersagemodell der Intelligenz liegen – wenn überhaupt – nur sehr wenigen Technologieunternehmen vor und können in der Regel nicht genutzt werden. Für die Forschung und Praxis wäre ein solcher Ansatz folglich nur umsetzbar, wenn es entweder eine zentrale Sammelstelle für die Daten gäbe, auf die zugegriffen werden kann, oder wenn die Versuchspersonen und Bewerbenden auf irgendeine Art und Weise selbst ihre Daten zur Verfügung stellen würden. Diese Überlegungen setzen jedoch eine ethische Grundlage voraus, die teilweise nur schwer erfüllbar ist. Da es sich um persönliche Daten handelt, müssen alle Teilnehmenden auch dazu in der Lage sein, eine freiwillige und gut informierte Entscheidung zu treffen, wann, mit wem und wie viele Daten sie teilen wollen. Für Forschungsvorhaben wäre es denkbar, dass Versuchspersonen einer Sammlung aller persönlicher Daten über einen bestimmten Zeitraum zustimmen. Daraus gewonnene Erkenntnisse könnten dann dazu führen, dass bessere Testverfahren entwickelt werden, die auch die Fairness für Bewerbende in der Praxis steigern. Denn besonders bei der Intelligenzdiagnostik in der Personalauswahl dürfte ein Sammeln von Daten schwer bis unmöglich sein. Entscheidet sich eine Person hier ihre Daten nicht zu teilen, so könnte ihr daraus ein (tatsächlicher oder nur wahrgenommener) Nachteil entstehen, der die Freiwilligkeit der Angabe reduziert.

Letztlich ist noch anzumerken, dass häufig angenommen wird, dass es sich bei künstlicher Intelligenz und Machine-Learning Modellen um eine Art Black Box handelt, bei der nicht ersichtlich ist, wie das Ergebnis zustande kommt. Dies ist jedoch nur begrenzt korrekt, da es sich letztendlich nur um lineare Kombinationen von Inputvariablen handelt, die keineswegs undurchdringlich sind (Kietzmann et al., 2019). Mithilfe der *network dissection* ist es möglich verschiedene Ebenen eines Machine-Learning Modells zu betrachten und zu analysieren, welche Inputvariablen zu einer maximalen Aktivierung innerhalb des Modells führen, und so zu schlussfolgern, welche

Variablen für weitere Forschung besonders interessant sind (Zhou et al., 2019). Damit können Ergebnisse aus Machine-Learning Modellen auch zur Theoriegenese genutzt werden und auch die Grundlagenforschung unterstützen.

Der dritte Ansatz zur Modernisierung der Intelligenzdiagnostik, der in Studie I (Koch, Becker, et al., 2021) besprochen wurde, bezieht sich auf die direkte Interaktion des Gehirns mit Computern durch BCIs. In der Chemie musste man früher zur Schätzung des pH-Wertes auf Lackmus Papier zurückgreifen, bis herausgefunden wurde, dass der pH-Wert durch Wasserstoffionen bestimmt wird und man ihn folglich mit Elektroden messen kann. Die Entwicklung von Intelligenztests kann man analog zum Lackmus Papier verstehen. Zwar wissen wir, dass intelligentes Verhalten durch das Zusammenspiel verschiedener Gehirnregionen zustande kommt und dass entsprechend neurowissenschaftliche Forschung zentral für das Verständnis von Intelligenz und damit Ihrer Messung ist (Haier, 1990), aber bisherige Forschung ist dadurch limitiert, dass die meisten BCIs sehr teuer und schlecht zugänglich sind (z.B. fMRT, EEG). Mit diesen Methoden hat man jedoch bereits herausgefunden, dass besonders parieto-frontale Regionen des Gehirns eine zentrale Rolle spielen (Jung & Haier, 2007) und dass bereits die Ruheaktivität des Gehirn mit der Intelligenz assoziiert ist (Hearne et al., 2016; Santarnecchi et al., 2017). Besonders zu erwähnen ist eine Studie, die einen Zusammenhang zwischen einem Ruhezustand fMRT und *g* von $r = .25$ und zu einem verbalen Intelligenzquotienten von $r = .54$ aufzeigen konnte (Dryburgh et al., 2020). Zwar ist dies nicht für eine Messung der Intelligenz ausreichend, kann aber doch als eindeutiger Anhaltspunkt verstanden werden, dass Intelligenz ohne Intelligenztests anhand neuronaler Strukturen und Aktivitäten messbar ist. Die Forschung ist hier jedoch noch durch technische und statistische Limitationen beschränkt. Je besser die Auflösung wird mit der Gehirnareale vermessen werden können, desto größer wird auch die Menge an Prädiktoren für Vorhersagemodelle. Klassische statistische Methoden weisen jedoch häufig einen Bias bei einer großen Prädiktorenanzahl auf (Lavery et al., 2019). Machine-Learning Modelle bieten hier einen alternativen Ansatz zur Auswertung, der inhärent auch durch Kreuzvalidierung gekennzeichnet ist und somit stabiler gegen Stichprobenfehler ist.

Auch dieser Forschungsansatz muss von gründlichen ethischen Diskussionen begleitet werden, da die Messung der Intelligenz am Gehirn in die Privatsphäre des Menschen eingreift. Darüber hinaus ist selbst eine zu 100% akkurate Messung der Intelligenz nicht ausreichend, um perfekte Vorhersagen für Kriterien treffen zu können. Zwar korreliert Intelligenz mit zahlreichen Variablen des täglichen Lebens (Gottfredson, 1997, 2004; Neisser et al., 1996), ist jedoch nie die alleinige Ursache. Trotzdem sollte dieser neurowissenschaftliche Ansatz ausführlich untersucht werden, da er einerseits das theoretische Verständnis von Intelligenz vertiefen kann, andererseits aber auch die Möglichkeit zur Diagnose bei bestimmten Personengruppen eröffnet, die man auf herkömmliche Art und Weise nicht testen kann (z.B. Säuglinge und Patienten mit locked-in Syndrom).

Diskussion

Studie I (Koch, Becker, et al., 2021) hat drei verschiedene Ansätze untersucht, die zu einer Modernisierung und Verbesserung der Intelligenzdiagnostik führen können. Während es zur Umsetzung des ersten Ansatzes – einer stärkeren Nutzung der Möglichkeiten computerbasierter Testung – nur einen teilweisen Wandel der Forschungsziele benötigt, so fehlen für den zweiten Ansatz – die Nutzung künstlicher Intelligenz zur Schätzung der Intelligenz – noch die notwendigen Daten. Für den dritten Ansatz müssen noch bestehende Geräte und statistische Methoden weiterentwickelt werden. Nichtsdestotrotz konnte für alle drei Bereiche aufgezeigt werden, dass es bereits erste Ergebnisse gibt, die in mittelbarer Zukunft eine Alternative zu aufwändigen Intelligenztestbatterien bieten könnten.

Für die Intelligenzforschung und die Intelligenzdiagnostik stellen diese Ansätze jedoch einige Herausforderungen dar. Es muss zwangsläufig zu einem stärkeren interdisziplinären Austausch zwischen Psychologie, Neurowissenschaft, Medizin und Informationstechnologie kommen, um alle Ansätze realisieren zu können und mit den neuen Anforderungen an Forschungsdaten umgehen zu können. Darüber hinaus stellt sich die Frage, ob Intelligenzdiagnostik eventuell obsolet wird, wenn eine Messung der Intelligenz am Gehirn erfolgen kann. Hierbei ist jedoch anzumerken, dass solide Theorien zur Messung der Intelligenz eine zentrale Voraussetzung sind, um

diese mit neueren Verfahren messen zu können. Aus diesem Forschungszweig sind also besonders Ansätze wie das CDS wertvoll, die konkrete Hypothesen über den Zusammenhang von Intelligenz(-aspekten) zu bestimmten Eigenschaften der Items aufstellen (Embretson, 1998) und es damit ermöglichen gezielt nach neuen technologischen Möglichkeiten zu suchen, um diese Aspekte auf andere Art und Weise zu erfassen.

Viele der bisherigen Ansätze gehen auch noch mit einem deutlich größeren Aufwand einher als traditionelle Intelligenztests und sind daher zumindest für die Praxis noch unattraktiv. Ein Zusammenhang von r = .79 zwischen Videospieleleistung und Intelligenz (Quiroga et al., 2019) ist zwar beeindruckend, jedoch mussten die Versuchspersonen dafür an zehn Videospielen teilnehmen (Bearbeitungszeit: 3 Stunden). Ähnlich ist eine fMRT Sitzung mit hohem finanziellem und organisatorischem Aufwand assoziiert und daher wenig ökonomisch, sodass die bisherigen Ergebnisse für die Praxis nur von geringer Relevanz sind.

Damit stellen die hier vorgestellten Ansätze also keinesfalls den Endpunkt der psychologische Intelligenztestentwicklung dar, sondern sollten vielmehr als Anreiz verstanden werden, neuartige Testverfahren zu entwickeln, die gleichzeitig höhere Validität besitzen, in kürzerer Zeit durchgeführt werden können und im Sinne einer besseren Zumutbarkeit weniger Anforderungen an die Teilnehmenden an solchen Verfahren stellen.

Zusammenfassung

Ziel von Studie I (Koch, Becker, et al., 2021) war es anhand bereits existierender Technologien und Forschung neue Wege aufzuzeigen, wie die Intelligenzdiagnostik modernisiert und weiterentwickelt werden kann. Dabei wurde festgestellt, dass besonders im Bereich des computerisierten Testens das Potential besteht, um eine Intelligenzdiagnostik zu betreiben, die neue Möglichkeiten eröffnet und nicht auf traditionelle Testformate festgelegt ist. Erste Studien mit neuen Ansätzen (Primi, 2014; Quiroga et al., 2019; Simons et al., 2021) existieren bereits und berichten vielversprechende Ergebnisse. Die Bereiche der künstlichen Intelligenz und der BCIs sind zur Zeit noch weniger weit entwickelt, Befunde aus der Persönlichkeitsforschung (Azucar et

al., 2018; Settanni et al., 2018) sowie Versuche Intelligenz mit fMRT zu erfassen (Dryburgh et al., 2020), deuten jedoch auf eine zentrale Rolle dieser Gebiete in der zukünftigen Intelligenzforschung hin. Studie I hat somit einen Ausgangspunkt für neue Forschungsvorhaben und Diskussionen im Bereich der psychologischen Leistungsdiagnostik geschaffen.

3 Studierendenauswahlverfahren

3.1 Allgemeine Einordnung

Laut Angaben des Statistischen Bundesamtes sind in Deutschland etwa 43% der Studiengänge zulassungsbeschränkt (Hochschulrektorenkonferenz, 2021), was die Universitäten vor die Herausforderung stellt, einen möglichst fairen und effektiven Studierendenauswahlprozess zu gestalten. Hierbei wurde lange Zeit ausschließlich die Abiturdurchschnittsnote als Kriterium genutzt. Jedoch wurde dies vom Bundesverfassungsgericht als teilweise verfassungswidrig erklärt (Bundesverfassungsgericht, 2017), da die Abiturnote auch stark vom Bundesland und dem soziökonomischen Status (SES) der Familie abhängt. Stattdessen wurde die Forderung nach einem fairen, für alle Bewerbenden zugänglichen Testverfahren gestellt.

Wie bereits in Kapitel 2 dargelegt wurde, weißt die Intelligenz eine hohe prädiktive Validität für akademischen Erfolg (Roth et al., 2015) und beruflichen Erfolg (Schmidt & Hunter, 2016) auf. Daraus ergibt sich, dass ein faires und valides Studierendenauswahlverfahren zumindest anteilsweise auch aus Intelligenztests bestehen sollte. Da ein Studierendenauswahlverfahren jedoch auch andere Fähigkeits- und Eigenschaftsbereiche erfassen muss (z.B. fachliche Kompetenzen, naturwissenschaftliche Grundkenntnisse, persönliche Passung etc.) sollte ein möglichst konziser Intelligenzindikator genutzt werden. Figurale Matrizenaufgaben sind eine Klasse von Intelligenztests, die besonders hoch auf fluider Intelligenz (Gignac, 2015) und damit auch auf *g* laden (Carpenter et al., 1990; Jensen, 1998).

Für die Studierendenauswahl ist es zentral vor allem auch im Hochleistungsbereich gut differenzieren zu können, daher empfiehlt sich der Einsatz einer figuralen Matrizenaufgabe, die auf dem Konstruktionsprinzip basiert und somit das Raten als Lösungsprozess eliminiert und damit die Konstruktvalidität des Testverfahrens verbessert (Becker et al., 2016).

Aus der Notwendigkeit von Studierendenauswahlverfahren und der damit einhergehend Wichtigkeit des Ergebnisses dieser Testungen, ergibt sich auch eine hohe Motivation von Studieninteressierten sich vorzubereiten. Diese Motivation dient als idealer Nährboden für „Testschulen“, die eine kommerzielle Vorbereitung auf

Studierendenauswahlverfahren anbieten, ähnlich wie es auch zahlreiche kostenpflichtige Vorbereitungsprogramme zur Medizinisch-Psychologischen Untersuchung gibt. In Studie II (Levacher et al., 2021) wurde untersucht, wie stark sich der Einfluss von unterschiedlicher Testvorbereitung auf die Testergebnisse in Studierendenauswahlverfahren auswirkt.

3.2 Einflüsse minimaler Interventionen auf Testfairness

Hinführung

Besonders in Situationen, in denen ein Testergebnisse starke Auswirkungen auf das persönliche Leben von Individuen hat, ist Testvorbereitung eine wichtige Einflussvariable, die untersucht werden muss (Buchmann et al., 2010). Kritisch ist vor allem, dass der Zugang zu Vorbereitungsangeboten vom SES der Familie abhängig ist und somit zu sozialer Ungerechtigkeit beiträgt. Generell ist eine grobe Unterteilung in drei Aspekte der Testvorbereitung möglich. Eine Möglichkeit besteht in der schlichten Wiederholung von Testaufgaben, was metaanalytisch betrachtet einen mittleren Effekt von d = 0.37 verursacht (Scharfen et al., 2018). Die zweite Möglichkeit besteht im Familiarisieren mit den Testaufgaben (z.B. in Form von Parallelversionen), wobei die Effekte etwas geringer ausfallen ($0.21 \leq d \leq 0.23$; Hausknecht et al., 2007; Kulik et al., 1984; Scharfen et al., 2018). Beim Testcoaching wird schließlich noch zusätzliches Wissen über die Testaufgaben (z.B. Themenbereiche, Fragentypen und Lösungsstrategien) vermittelt, was mit den größten Effekten assoziiert ist (d = 0.43; Kulik et al., 1984) und in Kombination mit Übung besonders effektiv ist (d = 0.64; Hausknecht et al., 2007).

Studien zur Analyse von Testvorbereitung bei figuralen Matrizen nutzen als Intervention häufig eine Erklärung der Konstruktionsprinzipien der Aufgaben und berichten dabei mittlere bis große Mittelwertsunterschiede zwischen Experimental- und Kontrollgruppe (Loesche et al., 2015; Schneider et al., 2020), die aber nicht mit einer Veränderung der Intelligenz assoziiert sind (Haier, 2014; Hayes et al., 2015).

Ziel der Studie

Ziel von Studie II (Levacher et al., 2021) war es, herauszufinden wie sich eine minimale Intervention direkt vor der Durchführung eines konstruktionsbasierten Matrizentests auf die resultierenden Testwerte auswirkt. Zusätzlich sollte die Faktorenstruktur zwischen der Experimental- und Kontrollgruppe verglichen werden, um zu untersuchen, inwiefern sich eine Veränderung der Ergebnisrohwerte auf die Faktorstruktur und damit letztendlich auch auf die Konstruktvalidität eines Tests auswirkt.

Methodik

Im Rahmen der Studierendenauswahltestung für das Uniklinikum Hamburg-Eppendorf nahmen *N* = 882 Personen (71.09% weiblich) teil, die sich 2018 für ein Medizinstudium beworben haben. Die Teilnehmenden waren im Durchschnitt 19.67 Jahre alt (*SD* = 2.01). Die Teilnahme an der figuralen Matrizenaufgabe war freiwillig, erfolgte am Ende der eigentlichen Studierendenauswahltestung und die Ergebnisse des Matrizentests wurden nicht für die Studienplatzvergabe genutzt. Die eigentlich Studierendenauswahl wurden anhand eines Naturwissenschaftstests (HamNat) durchgeführt.

Vor Beginn der Testung wurden die Teilnehmenden zufällig in zwei Gruppen aufgeteilt, sodass *n* = 461 Personen (69.63% weiblich; M_{Alter} = 19.70, *SD* =2.08) der Experimentalgruppe und *n* = 421 Personen (72.68% weiblich; M_{Alter} = 19.63, *SD* =1.93) der Kontrollgruppe zugewiesen wurden. Beide Gruppen erhielten eine DIN-A4 mit schriftlichen Instruktionen bezüglich der Testbearbeitung, in der erklärt wurde, wie das konstruktionsbasierte Antwortformat der figuralen Matrizen zu bearbeiten ist. Darüber hinaus erhielt die Experimentalgruppe eine weitere DIN-A4 Seite mit einer schriftlichen Erläuterung der Konstruktionsregeln, die zur Item Generierung genutzt wurden (z.B. „Addition: Die Elemente der ersten und der zweiten Zelle werden kombiniert. In der dritten Zelle sind daher beide Elemente enthalten.“). Die Teilnehmenden hatten zehn Minuten Zeit, sich mit den schriftlichen Instruktionen vertraut zu machen, und hatte im Anschluss 20 Minuten zur Bearbeitung von 28 Aufgaben aus dem DESIGMA (Becker & Spinath, 2014).

Um einen möglichst messfehlerfreien Schätzer für die Mittelwertsdifferenz der beiden Gruppen zu erhalten, wurde ein Strukturgleichungsmodell mit Hilfe der Statistiksoftware R (R Core Team, 2021) sowie dem Paket *lavaan* (Rosseel, 2012) geschätzt. Dies erfolgte, indem der Mittelwert der Kontrollgruppe auf Null fixiert wurde, während der Mittelwert der Experimentalgruppe frei variieren durfte (Finch & French, 2015). Zum Vergleich der Faktorenstruktur in beiden Bedingungen wurde eine Reihe von vier konfirmatorischen Multigruppen-Faktoranalysen (MGCFA; multigroup confirmatory factor analysis) geschätzt. In der ersten Analyse wurde lediglich die Faktorstruktur konstant gehalten (konfigurales Modell). In der zweiten Analyse wurden darüber hinaus die Ladungen zwischen den Aufgaben und *g* konstant gehalten (schwaches Invarianzmodell). In der dritten Analyse wurde zusätzlich zu den Ladungen auch die Intercepts in beiden Gruppen auf den gleichen Wert fixiert (starkes Invarianzmodell). In der vierten Analyse wurden schließlich auch die Residuen konstant gehalten (strenges Invarianzmodell). Um auch Unterschiede auf Itemebene untersuchen zu können wurde darüber hinaus eine Analyse für differentielle Itemfunktion (DIF) berechnet (Penfield & Camilli, 2006).

Um zu untersuchen inwiefern sich die Intervention auf die Kriteriumsvalidität der figuralen Matrizenaufgabe auswirkt, wurde außerdem für beide Gruppen eine bivariate Korrelation zwischen dem Summenscore der figuralen Matrizen sowie dem Ham-Nat berechnet und die Differenz der Korrelationen auf Signifikanz getestet (Diedenhofen & Musch, 2015).

Ergebnisse

Während in der Kontrollgruppe nur 41% (*SD* = 0.25) der Aufgaben korrekt gelöst wurden, so konnten die Teilnehmenden der Experimentalgruppe im Durchschnitt 61% (*SD* = 0.19) richtig lösen. Der Test auf Differenz der latenten Mittelwerte wurde signifikant (ΔM = 1.48, *SD* = 1.49, 95% CI = [1.25;1.71]; χ^2 = 154.09, $p < .001$). Dies ist gleichbedeutend mit einem großen Effekt zugunsten der Experimentalgruppe (*d* = 0.94, 95% CI = [0.80;1.08]). Die Korrelation zwischen den Itemschwierigkeiten in beiden Gruppen war dennoch sehr hoch ($r = .99$, $p < .001$).

Die Ergebnisse der MGCFAs sind in Tabelle 1 abgetragen. Der Modellfit ist, bis einschließlich zum starken Invarianzmodell, als gut zu bewerten (CFI > .95, RMSEA < .06; Hu & Bentler, 1999), es kann also angenommen werden, dass unabhängig vom Erlernen der Konstruktionsregeln die Faktorenstruktur und der Zusammenhang der einzelnen Items mit *g* bestehen bleibt. Das Scheitern des strikten Invarianzmodells ist konsistent zum berichteten Unterschied in den latenten Mittelwerten.

Tabelle 1. *Ergebnisse der MGCFAs zum Vergleich der Faktorstruktur des Matrizentests zwischen beiden Gruppen.*

Modell	χ2	*df*	*p*(χ2)	CFI	ΔCFI	RMSEA	ΔRMSEA	Δχ2	*Δdf*	*p*(Δχ2)
Konfigural	3.21	4	.52	1	-	<.001	-	-	-	-
Schwach	12.22	7	.09	.99	.001	.04	.04	9.01	3	.03
Stark	22.96	10	.01	.99	.002	.05	.01	10.75	3	.01
Streng	177.05	11	<.001	.94	.05	.19	.14	154.09	1	<.001

Notiz: χ2 = Teststatistik des Modellfits; *df* = Freiheitsgrad; *p*(χ2) = Wahrscheinlichkeit der Teststatistik unter Annahme der Nullhypothese; CFI = Comparative Fit Index; RMSEA = Residual Mean Square Error of Approximation; Δ = Veränderung zum jeweils vorherigen Modell.

Die DIF Analyse ergab anhand des Waldtests 13 Items mit signifikanten Unterschieden zwischen den beiden Bedingungen, während anhand der Mantel-Haenszel Statistik nur für sechs Items eine signifikante DIF festgestellt wurde. Davon lag bei zwei Items eine starke DIF vor, bei den übrigen vier Items eine moderate. Die genaue Itemparameter und DIF Statistiken sind in Tabelle 2 dargestellt.

Ein Vergleich der Kriteriumsvalidität zwischen beiden Gruppen ergab keinen signifikanten Unterschied (z = -0.71, p = .48) der Korrelation des Matrizentests mit einem Naturwissenschaftstest zwischen der Experimentalgruppe (r = .28, p < .001) und der Kontrollgruppe (r = .24, p < .001).

Tabelle 2. *Itemparameter und DIF Statistiken.*

Item	Rules	p_p	p_c	r_p	r_c	b_p	b_c	*Wald*	$Wald_p$	MH_Δ	MH_e
1	1	.94	.95	.31	.20	-2.35	-2.52	3.50	< .001	-1.79	C
2	1	.92	.90	.40	.34	-2.06	-1.92	2.79	.01	-1.25	B
3	1	.95	.93	.33	.24	-2.44	-2.29	1.95	.05	-0.69	A
4	1	.83	.80	.17	.12	-1.39	-1.26	4.65	< .001	-0.03	A
5	1	.84	.58	.45	.37	-1.41	-0.33	-2.70	.01	1.65	C
6	2	.63	.43	.43	.40	-0.47	0.26	1.46	.15	0.14	A
7	2	.77	.49	.45	.48	-1.04	0.03	-2.35	.02	1.46	B
8	2	.55	.32	.56	.43	-0.19	0.68	0.40	.69	-0.08	A
9	2	.67	.40	.57	.49	-0.60	0.36	-0.85	.39	0.29	A
10	2	.77	.51	.51	.58	-1.05	-0.06	-1.60	.11	0.70	A
11	2	.63	.46	.53	.36	-0.45	0.13	2.89	< .001	-0.62	A
12	2	.78	.63	.54	.47	-1.12	-0.48	1.31	.19	-0.50	A
13	2	.44	.21	.48	.40	0.20	1.18	-0.25	.81	0.23	A
14	3	.66	.51	.53	.54	-0.58	-0.06	3.33	< .001	-1.34	B
15	3	.60	.38	.59	.59	-0.36	0.45	0.85	.39	-0.86	A
16	3	.64	.37	.60	.58	-0.50	0.47	-0.84	.40	-0.08	A
17	3	.65	.33	.72	.62	-0.54	0.66	-2.89	< .001	0.15	A
18	3	.61	.42	.54	.54	-0.40	0.29	1.89	.06	-0.84	A
19	3	.57	.39	.51	.52	-0.25	0.41	2.36	.02	-.083	A
20	3	.60	.23	.65	.57	-0.36	1.07	-4.70	< .001	1.45	B
21	3	.65	.36	.61	.63	-0.53	0.51	-1.52	.13	0.04	A
22	4	.39	.15	.52	.44	0.36	1.52	-1.52	.13	0.30	A
23	4	.45	.19	.61	.58	0.17	1.33	-1.75	.08	-0.45	A
24	4	.48	.20	.61	.49	0.08	1.28	-2.14	.03	0.23	A
25	4	.35	.11	.55	.51	0.53	1.83	-2.41	.02	0.22	A
26	5	.30	.11	.49	.46	0.73	1.90	-1.20	.23	-0.15	A
27	5	.35	.10	.56	.40	0.53	1.92	-2.96	< .001	0.72	A
28	5	.24	.07	.47	.45	0.97	2.29	-1.90	.06	-0.10	A

Notiz: p_P = Lösungswahrscheinlichkeit des Items in der Experimentalgruppe; p_C = Lösungswahrscheinlichkeit des Items in der Kontrollgruppe; r_P = Trennschärfe des Items in der Experimentalgruppe; r_C = Trennschärfe des Items in der Kontrollgruppe; b_P = Rasch-modellierte Itemschwierigkeit in der Experimentalgruppe; b_C = Rasch-modellierte Itemschwierigkeit in der Kontrollgruppe; Wald = Waldtest Teststatistik; $Wald_p$ = Wahrscheinlichkeit der Teststatistik unter Annahme der Nullhypothese; MH_Δ = Mantel-Haenszel Differenz; MH_e = Klassifikation von MH_Δ; A = Keine DIF; B = Moderate DIF; C = Große DIF.

Diskussion

Studie II (Levacher et al., 2021) versuchte im Vergleich zu vorherigen Studien (Losche et al., 2015; Schneider et al., 2020) herauszufinden, wie stark sich ein minimales Training in Form schriftlicher Informationen auf die Fairness und Validität eines figuralen Matrizentests auswirkt.

Wie erwartet wurde ein großer Effekt für die Differenz in den gelösten Items zwischen den beiden Gruppen gefunden, der sich in einem vergleichbaren Bereich anordnet wie auch die bisherigen Befunde (Loesche et al., 2015; Schneider et al., 2020). Im Vergleich zu den bisherigen Studien ist die Intervention der vorliegenden Studie jedoch deutlich ökonomischer in der Durchführung. Die Vergleichbarkeit der Effektstärken deutet also daraufhin, dass das Wissen über die Konstruktionsprinzipien figuraler Matrizen sehr schnell vermittelt werden kann und einen zentralen Faktor für die Lösungswahrscheinlichkeit darstellt.

Darüber hinaus konnte gezeigt werden, dass der figuralen Matrizentest sich weder in seiner Struktur verändert in Abhängigkeit davon, ob die Teilnehmenden Regelkenntnis besitzen oder nicht, noch Unterschiede in der Korrelation zu einem externen Kriterium aufweist. Diese Ergebnisse sprechen dafür, dass figurale Matrizenaufgaben auch nach dem Erlernen der Konstruktionsregeln weiterhin valide Intelligenztestverfahren bleiben und zur Differenzierung zwischen leistungsfähigen Personen eingesetzt werden können. Die Befunde können auch derart interpretiert werden, dass Regelkenntnisse lediglich eine lineare Verschiebung der Itemschwierigkeitsparameter verursachen, Itemschwierigkeit und Personenfähigkeit jedoch weiterhin auf der gleichen Dimension verankert sind. Die Fähigkeit die Einzelregeln korrekt zu erkennen, von anderen Regeln zu differenzieren und in Kombination mehrerer Regeln korrekt zur Konstruktion der Lösung zu nutzen, ist somit weiterhin zentral für die Bearbeitung der figuralen Matrizen.

Während der Waldtest zwar 13 Items mit signifikanter DIF identifizierte, wurden anhand der Mantel-Haenszel Statistik nur sechs Items signifikant. Die Mantel-Haenszel Statistik ist insgesamt robuster und lässt eine Einschätzung der Effektgröße zu (Wetzel & Böhnke, 2017). Eine Analyse der sechs Items mit signifikanter DIF ergab keine systematischen Gemeinsamkeiten, sodass keine Aussage getroffen werden kann, woher diese Unterschiede stammen könnten.

Um einen möglichst fairen Einsatz von figuralen Matrizenaufgaben zur Studierendenauswahl gewährleisten zu können empfiehlt sich also allen Teilnehmenden eine Übersicht und Erklärung der verwendeten Konstruktionsprinzipien anzubieten. Anhand der aktuellen Ergebnisse lässt sich noch nicht beurteilen, wie wirksam diese

Intervention ist, wenn ein größerer zeitlicher Abstand zwischen Training und Testung liegt. Da es sich jedoch um eine minimale Intervention handelt, können aber zusätzlich zu der normalen Instruktion immer auch die Konstruktionsprinzipien in schriftlicher Form dargeboten werden, um allen Teilnehmenden die gleichen Ausgangsbedingungen zu ermöglichen. Folgende Studien sollten zusätzlich noch evaluieren, ob ein zusätzliches Training (z.B. Videos oder Trainingsaufgaben mit Feedback) noch zu einer weiteren Verbesserung der Teilnehmenden führt, oder ob durch die Regelkenntnis bereits eine Sättigung des Trainings erreicht wurde.

Zusammenfassung

Studie II (Levacher et al., 2021) ergänzt die bestehende Literatur zur Trainierbarkeit von figuralen Matrizenaufgaben durch den Einsatz einer ökonomischeren Intervention und einer Analyse der Konstruktvalidität. Die Ergebnisse sprechen dafür, dass die Validität figuraler Matrizen nicht durch ein vorheriges Vertrautmachen mit den Konstruktionsprinzipien beeinträchtigt wird und folglich für die Studierendenauswahl eingesetzt werden können. Um einen möglichst fairen Prozess zu ermöglichen, sollte allen Teilnehmenden ein kostenfreier Zugang zu den Vorbereitungsmaterialien ermöglicht werden.

4 Computerbasierte (Intelligenz-)Diagnostik

4.1 Allgemeine Einordnung

Computerbasierte Diagnostik beschreibt den Einsatz von Computersystemen zur Durchführung und Auswertung von psychologischen Testverfahren. Heutzutage besteht großer Konsens darüber, das computerbasierte Diagnostik viele Vorteile gegenüber herkömmlichen Papiertestungen bietet (Tippins, 2015). Dazu zählen unter anderem die Einfachheit und Fehlerfreiheit von Auswertungsprozeduren, zugeschnittene Ergebnisrückmeldungen, uneingeschränkte Objektivität sowie die Möglichkeit zur Erfassung von weiteren Prozessdaten wie etwa Reaktionszeiten. Darüber hinaus gibt es sogar einzelne Befunde, dass computerbasiertes Testen zur Reduktion eines Bias gegen ethnische Minderheiten führen kann im Vergleich zu papierbasierten Testverfahren (Gallagher et al., 2002).

Obwohl der Zugang zu Computern früher noch deutlich weniger verbreitet war als heute und die Anschaffung auch mit deutlich höheren Kosten verbunden war, finden sich schon sehr frühe Versuche der Konversion von papierbasierten zu computerbasierten Verfahren (Sachar & Fletcher, 1978). Mittlerweile sind Computer dank der rasanten technologischen Fortschritte der vergangenen Jahrzehnte zu einer selbstverständlichen Grundausstattung psychologischer Labore geworden und finden sich in der industrialisierten Welt auch in fast allen Haushalten wieder. Metaanalytische Befunde sprechen mittlerweile auch deutlich dafür, dass die Ergebnisse psychologischer Leistungstests auf beiden Darbietungsmodalitäten vergleichbar sind (Mead & Drasgow, 1993; Wang et al., 2007, 2008). Lediglich bei Testverfahren mit einer starken Speed-Komponente gibt es metaanalytisch signifikante Unterschiede, die zu einem Nachteil bei der Teilnahme an computerbasierten Testungen gegenüber herkömmlichen papierbasierten Testungen führen.

Es steht außer Frage, dass eine Fortentwicklung der psychologischen Intelligenzdiagnostik daher auf einer stärkeren Nutzung von computerbasierter Diagnostik beruht. Einen besonderen Stellenwert nimmt an dieser Stelle das *item banking* und *tailored testing* ein. Beim *tailored testing* handelt es sich um eine bestimmte Vorgehensweise bei der Diagnostik, in der der zu testenden Person nur eine Auswahl aller möglichen

Items vorgelegt wird (van der Linden & Pashley, 2009). So wird in der Regel mit einem mittelschweren Item gestartet, bei richtiger Antwort folgt ein schwereres, bei falscher Antwort ein leichteres Item. Diesen Prozess setzt man so lange fort, bis das maximale Leistungsniveau der Testperson erreicht ist. Während das *item banking* zwar keine absolute Voraussetzung für *tailored testing* ist, so vereinfacht es den Prozess dennoch sehr stark. So kann jede ausreichend große Itembank zur Entwicklung eines Testverfahrens genutzt werden, dass auf *tailored testing* basiert. Beim *item banking* handelt es sich um eine Strategie, einen großen Item Pool zu generieren, aus dem eine für einen bestimmten Zeck notwendige Menge an Items extrahiert wird und in beliebiger Kombination zu einem Test verknüpft wird (Weiss, 2013). Dabei ist die Item Response Theorie (IRT) die methodische Grundlage für das *item banking.* In der IRT wird postuliert, dass Itemparameter (also die Schwierigkeit eines Items) und Personenparameter (also die Fähigkeit einer Person Aufgaben eines bestimmten Typs zu lösen) auf der gleichen Dimension verortet sind. Weisen alle Items unter den Annahmen des IRT Modells einen guten Modellfit auf, so kann angenommen werden, dass jedes einzelne Item das gleiche zugrundeliegende Konstrukt misst und somit in beliebiger Kombination mit anderen Items dargeboten werden kann.

4.2 Entwicklung und Validierung von Itembanken

Hinführung

Bei einer Itembank handelt es sich um eine Sammlung von Items zur Messung eines spezifischen Konstrukts (Chituc et al., 2019; Ward & Murray-Ward, 1994). In der Regel werden zur Erstellung einer Itembank auch IRT Modelle genutzt, um zu gewährleisten, dass alle Items das gewünschte Konstrukt auf der gleichen Dimension abtragen und somit auch in beliebigen Kombinationen vereinbar sind (Bjorner et al., 2007).

Ziel der Studie

Ziel von Studie III (Koch, Spinath, et al., 2021) war die Entwicklung einer Itembank für konstruktionsbasierte figurale Matrizenaufgaben zur freien Verwendung in der

wissenschaftlichen Forschung. Zu diesem Zweck sollten 220 Items generiert und psychometrisch evaluiert werden, die einen möglichst breiten Fähigkeitsbereich abdecken.

Methodik

Um ein homogenes Aussehen sowie eine durchgehend hohe Qualität der generierten Items zu garantieren, wurde in einem ersten Schritt eine neue Software zur technikgestützten Entwicklung figuraler Matrizen generiert. Diese Software ermöglicht das Generieren von Items durch einfaches Anklicken von geometrischen Elementen, bietet Regeldefinitionen für den Anwender an und prüft die Items auf Lösbarkeit.

Insgesamt wurden sechs Konstruktionsregeln in verschiedenen Kombinationen implementiert (Becker et al., 2016). (1) Addition – Elemente einer Reihe werden kombiniert, sodass in der letzten Zelle alle Teilelemente vorhanden sind. (2) Subtraktion – Die Elemente der zweiten Zelle werden von der ersten Zelle subtrahiert. (3) Einzelkomponentenaddition – Es werden in der letzten Zelle nur die Elemente abgebildet, die in den vorherigen Zellen einzigartig waren. (4) – Schnittmenge – In der letzten Zelle werden die gemeinsamen Elemente der vorherigen Zellen abgebildet. (5) Rotation – Die Elemente rotieren über die Zellen hinweg im oder gegen den Uhrzeigersinn. (6) Vollständigkeit – Eine bestimmte Kombination muss über alle Reihen hinweg gleichmäßig vorhanden sein.

Da bereits bekannt ist, dass die Schwierigkeit figuraler Matrizen zum größten Teil durch die Anzahl verwendeter Konstruktionsregeln bestimmt wird (Becker et al., 2016; Primi, 2014) wurden die Items für Studie III (Koch, Spinath, et al., 2021) so konstruiert, dass mehr mittelschwere als schwere und leichte Items genutzt werden können. Eine genau Verteilung der Konstruktionsregeln ist in Tabelle 3 abgebildet. Insgesamt wurden mit Hilfe der neu entwickelten Software 220 figurale Matrizen von zwei Bachelor-Studierenden generiert.

Tabelle 3. *Verteilung der Konstruktionsregeln in den neu konstruierten Items.*

Regeln	Anzahl Items	ADD	SUB	EKA	SM	ROT	VOL
Eine	20	4	4	3	3	3	3
Zwei	50	20	16	16	16	16	16
Drei	80	44	44	37	41	37	37
Vier	50	35	36	32	35	31	31
Fünf	20	17	17	17	17	16	16
Summe	220	120	117	105	112	103	103

Notiz: ADD = Addition; SUB = Subtraktion; EKA = Einzelkomponentenaddition; SM = Schnittmenge; ROT = Rotation; VOL = Vollständigkeit.

Zur Evaluation der neu konstruierten Aufgaben wurden im Rahmen einer Vorbereitungsstudie zur Studierendenauswahl *N* = 4657 Personen in Form einer selbstadministrierten Onlineerhebung getestet. Den Teilnehmenden wurden insgesamt acht verschiedene Testverfahren in randomisierter Reihenfolge dargeboten, mit denen sie sich auf die kommenden Studierendenauswahltestungen kostenlos vorbereiten konnten. Nachdem 2096 Datensätze aufgrund fehlender Teilnahme an der Matrizenaufgabe ausgeschlossen werden mussten, wurden für alle Analysen die Daten von *N* = 2561 Teilnehmenden genutzt.

Um die Belastung für jedes Individuum zu minimieren, wurde ein Item-Anchoring Design implementiert. Dazu wurden 10 Itemsets mit jeweils 22 Items generiert und um sechs Ankeritems erweitert. Diese Ankeritems waren über alle Itemsets hinweg identisch und konnten somit zur Nivellierung von etwaigen Stichprobenunterschieden genutzt werden (Battauz, 2017). Bei Umfragestart wurde jede Person randomisiert einem der 10 Itemsets zugewiesen und erhielt anhand von zwei Beispielaufgaben, die Möglichkeit das Antwortformat (s. Abb. 1) einzuüben. Nach erfolgreicher Bearbeitung der beiden Beispielaufgaben wurde den Teilnehmenden mitgeteilt, dass sie nun 30 Minuten Zeit zur Bearbeitung der Aufgaben haben. Nach Abschluss der Testung erhielten alle Teilnehmenden ein individualisiertes Feedback, in dem ihre eigenen Punktwerte im Vergleich zur Gesamtstichprobe dargestellt und erläutert wurden.

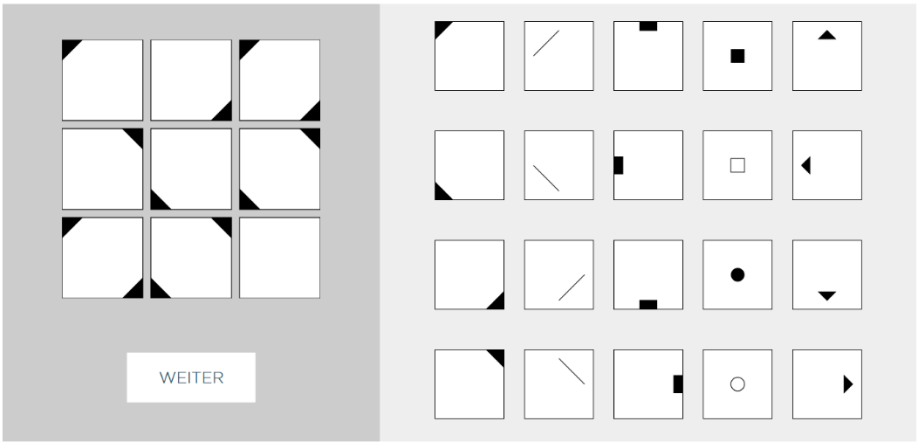

Abbildung 1*. Darstellung des Antwortformats der konstruktionsbasierten figuralen Matrizenaufgabe. Angeklickte Konstruktionselemente werden durch einen roten Rahmen markiert.*

Für alle Items wurden die Lösungsschwierigkeit und Trennschärfe entsprechend der klassischen Testtheorie mit dem Paket *psych* (Revelle, 2017) für die Statistiksoftware *R* (R Core Team, 2021) berechnet. Um Items zu identifizieren, wurden mit Hilfe des Pakets *mirt* (Chalmers, 2012) infit und outfit Statistiken extrahiert. Items wurden als unpassend zum 2-PL IRT Modell identifiziert, wenn eine Kombination aus einem infit oder outfit größer als 1.33 oder kleiner als 0.75 sowie eine signifikante t-Statistik vorlag (Wilson, 2005). Für alle akzeptierten Items wurden durch das Paket *equateMultiple* (Battauz, 2021) die Schwierigkeits- und Diskriminationsparameter geschätzt. Schließlich wurde mit einer Reihe von MGCFAs mit dem Paket *lavaan* (Rosseel, 2012) überprüft, ob alle 10 Itemsets die gleiche Faktorstruktur aufweisen.

Ergebnisse

Im Durchschnitt wurden von den Teilnehmenden 16.82 Items korrekt gelöst (*SD* = 7.79) und sie verbrachten durchschnittlich 26.05 Minuten mit der Aufgabe (*SD* = 7.07). Die Schwierigkeit der 10 Itemsets unterschied sich signifikant ($F_{(9,210)}$ = 4.66, p < .001, ω^2 = .13) ebenso wie die Trennschärfe ($F_{(9,210)}$ = 5.19, p < .001, ω^2 = .15). Die Mittelwerte und Standardabweichungen sind in Tabelle 4 dargestellt.

Tabelle 4. *Mittlere Itemschwierigkeiten und Trennschärfen der 10 Itemsets.*

Itemset	Schwierigkeit (*SD*)	Trennschärfe (*SD*)
1	.71 (0.15)	.44 (0.08)
2	.53 (0.16)	.58 (0.11)
3	.53 (0.18)	.54 (0.14)
4	.64 (0.14)	.51 (0.10)
5	.57 (0.16)	.57 (0.12)
6	.64 (0.16)	.52 (0.14)
7	.62 (0.16)	.52 (0.09)
8	.65 (0.10)	.59 (0.08)
9	.46 (0.19)	.55 (0.17)
10	.68 (0.22)	.41 (0.13)

Für die Schätzung der IRT basierten Itemparameter musste ein Item wegen schlechter Modellpassung ausgeschlossen werden. Die durchschnittlichen Itemparameter sind in Tabelle 5 dargestellt (eine Aufstellung auf Itemebene findet sich im Anhang der Studie III). Die durchschnittliche Schwierigkeit aller Items liegt bei b = -0.17 (*SD* = 0.99) und der durchschnittliche Itemdiskriminationsparameter beträgt a = 2.09 (*SD* = .84). Die IRT-basierte Itemschwierigkeit korreliert hoch mit der anhand der klassischen Testtheorie berechneten Itemschwierigkeit (r = -.68, p < .001) und diese Korrelation steigt bei einer Korrektur für extreme Ausreißer (± 3SD) stark an (r = -.83, p <.001). Des Weiteren wird die Itemschwierigkeit durch die Anzahl verwendeter Regeln vorhergesagt (r = .53, p <.001).

Tabelle 5. *IRT-basierte Itemparameter in Abhängigkeit von der Regelanzahl*

Regeln	a	Min_a	Max_a	b	Min_b	Max_b
Eine	1.45	0.11	3.16	-1.87	-8.98	1.43
Zwei	1.52	0.62	2.97	-0.30	-2.25	1.44
Drei	2.01	1.05	3.63	-0.16	-1.12	1.65
Vier	2.64	1.08	5.16	0.24	-0.12	0.88
Fünf	3.10	1.63	4.48	0.67	0.27	2.41
Durchschnitt	2.09	0.11	5.16	-0.17	-8.98	2.41

Note: a = Itemdiskriminationsparameter, b = Itemschwierigkeitsparameter, Min = Minimum, Max = Maximum.

Eine Analyse der Faktorstruktur unterstütze die Annahme der konfiguralen Invarianz zwischen allen 10 Itemsets (CFI = .991, RMSEA = .047, SRMR = .097). Ein stärkeres Invarianzmodell, das zusätzlich zur Faktorstruktur auch die Faktorladungen und Intercepts zwischen den Itemsets konstant hält, passt signifikant schlechter zu den Daten ($\Delta\chi^2$ = 448.96, Δdf = 234, p < .001).

Diskussion

Ziel von Studie III (Koch, Spinath, et al., 2021) war die Entwicklung einer Itembank für konstruktionsbasierte figurale Matrizenaufgaben. Hierzu wurden 220 Items generiert, von denen lediglich ein Item die Modellannahmen des 2-PL Modells verletzt. Durch das genutzt Item-Anchoring Design konnten für die übrigen 219 Items homogen skalierte Itemparameter geschätzt werden. Diese 219 Items dienen somit auch als Grundlage zur Zusammenstellung von Testverfahren mit einer vorher bestimmbaren Schwierigkeit.

Die mittlere Itemschwierigkeit der neu generierten Items liegt nahe bei $b = 0$, enthält aber auch sehr leichte und sehr schwere Items, sodass die Itembank auch zur Differenzierung in Extremgruppen genutzt werden kann. Während die minimale Itemschwierigkeit bei $b_{Min} = -8.98$ liegt, so beträgt die Schwierigkeit des schwersten Items $b_{Max} = 2.41$. Dieses Ungleichgewicht wird durch die Items verursacht, in denen nur eine Regel implementiert ist (z.B. Abb. 1). Betrachtet man nur die Schwierigkeiten aller Items mit mindestens zwei kombinierten Regeln, so ist die Schwierigkeitsverteilung deutlich symmetrischer ($b_{Min} = -2.25$). Für die Praxis bedeutet dies, dass Items, in denen nur eine Regel genutzt wird, entweder zur Verdeutlichung der Instruktionen genutzt werden sollten oder aber in besonders leistungsschwachen Stichproben eingesetzt werden sollten (z.B. sehr junge Kinder, Patienten mit neurologischen Erkrankungen). Werden noch schwerere Items benötigt, so können auch Items generiert werden, in denen gleichzeitig alle Regeln verwirklicht werden. Für die vorliegende Studie wurde eine Begrenzung auf fünf Regeln vorgenommen, da in früheren Studien mehr Regeln kaum mehr zur Differenzierung zwischen Individuen beigesteuert haben (Becker & Spinath, 2014).

Darüber hinaus wurde das Ergebnis früherer Studien bestätigt, dass die Anzahl genutzter Konstruktionsregeln ein starker Prädiktor für die Itemschwierigkeit ist (Becker et al., 2016; Primi, 2014). Sollen zusätzlich zu den bestehenden Items in der Itembank noch neue figurale Matrizen generiert werden, so können also bereits vor einem Einsatz im Feld erste Schätzer bezüglich der Itemparameter gewonnen werden.

Eine zentrale Limitation der vorliegenden Studie liegt in der Auswahl der Stichprobe. Obwohl die Teilnahme an der Vorbereitungsstudie freiwillig und kostenlos war, so

wurden dennoch nur Personen eingeladen, die sich für das anstehende Studierendenauswahlverfahren beworben hatten. Dies ist mit Kosten von etwa 100€ pro Person assoziiert, was zu einer gewissen Selbstselektion führen kann. Personen mit einer eher schlechten Abiturnote (und damit assoziiert wahrscheinlich geringerer Intelligenz; Roth et al., 2015) könnten von einer Bewerbung abgesehen haben, da sie ihre Chancen bei der Platzvergabe als zu gering einschätzen. Hierdurch war die Stichprobe in Studie III wahrscheinlich leistungsfähiger als die Allgemeinheit, was zu einer etwas verzerrten Schätzung der Itemparameter geführt haben könnte. Nichtsdestotrotz war nur etwa 1% der Teilnehmenden dazu in der Lage alle Items korrekt zu lösen, ein Deckeneffekt ist also ausgeschlossen. Darüber hinaus wurden die figuralen Matrizen aus der neu generierten Itembank auch bereits bei Schulkindern in einer Pilotstudie erfolgreich eingesetzt und haben sich bereits in anderen Studien etabliert (z.B. Krautter et al., 2021).

Zusammenfassung

Studie III (Koch, Spinath, et al., 2021) konnte eine für die Forschung kostenlose und frei zugängliche Itembank generieren, die auf den gleichen Konstruktionsprinzipien basiert wie bereits etablierte Matrizentests (Becker & Spinath, 2014). Aufgrund der für die Matrizenbank entwickelten Software ist eine Weiterentwicklung und Generierung von neuen Testaufgaben jederzeit möglich, sodass dieses Programm für diverse wissenschaftliche Projekt eingesetzt werden kann, ohne knapp bemessene Ressourcen zu benötigen.

5 Zusammenfassende Betrachtung

Aufgrund seiner zahlreichen Assoziation mit anderen Konstrukten (Gottfredson, 2004; Neisser et al., 1996; Roth et al., 2015; Schmidt & Hunter, 2016, 1998) ist Intelligenz eine der wichtigsten Variablen in der psychologischen Forschung. Aus diesem Grund existiert eine große Anzahl von Intelligenztestverfahren, die auf vergleichbaren Modellen, Annahmen und Konstruktionsprinzipien basieren. Aufgrund des technologischen Fortschritts (z.B. Smartphone, Wearables, VR) bieten sich immer wieder neue Möglichkeiten zur Untersuchung der Intelligenz, während gleichzeitig neue gesellschaftliche Anforderungen entstehen (z.B. Einsatz von Intelligenztests zur fairen Studierendenauswahl).

Das Ziel dieser Dissertation ist aus diesem zeitlichen Wandel entstanden und besteht in einer Modernisierung und Verbesserung der Intelligenzdiagnostik. Zu diesem Zweck wurde zunächst in Studie I (Koch, Becker, et al., 2021) analysiert inwiefern sich die Intelligenzdiagnostik in den letzten Jahrzehnten entwickelt hat und an welchen Stellen sie nicht mehr auf dem aktuellsten Stand der technologischen Möglichkeit ist. Obwohl bestehende Intelligenztests sich über lange Zeit etabliert haben und sie zu den psychologischen Testverfahren mit der höchsten Reliabilität und Validität gehören, wurden drei Entwicklungsbereiche identifiziert, in denen weitere Fortschritte möglich sind. Dazu zählt die Entwicklung computerbasierter Testverfahren, die nicht nur die Durchführung und Auswertung automatisieren sondern auch neue Datenströme eröffnen, die einer genaueren Untersuchung der Intelligenz dienen können (Dodonov & Dodonova, 2012; Greiff et al., 2015; Hunt & Pellegrino, 1985; Krieger et al., 2019). Dabei müssten diese eventuell gar nicht mehr den Anschein eines Intelligenztests erwecken, sondern könnten die Motivation der Teilnehmer durch einen spieleartigen Charakter steigern (Quiroga et al., 2017; Simons et al., 2021). Des Weiteren wurde thematisiert, dass ein Vorhersage der Intelligenz aufgrund stetig besser werdender Machine-Learning Algorithmen und einer größeren Datenmenge eine Alternative zu aufwändigen Leistungstestungen darstellen könnte (Azucar et al., 2018; Settanni et al., 2018). Schließlich wurde zusammengefasst, dass bereits heutzutage mit aktuellen bildgebenden Verfahren und durch sie messbaren physiologischen Variablen relativ genaue Schätzungen der menschlichen Intelligenz möglich sind

(Dryburgh et al., 2020; Hearne et al., 2016; Santarnecchi et al., 2017), was besonders für die Arbeit mit Stichproben wertvoll ist, die nicht in der Lage sind an herkömmlichen Testungen teilzunehmen.

Studie II (Levacher et al., 2021) beschäftigte sich mit der Testfairness von figuralen Matrizenaufgaben, wenn nur eine Hälfte der Teilnehmenden auf die Aufgaben vorbereitet wurde. Diese Fragestellung hat besonders hohe praktische und gesellschaftliche Relevanz, da figurale Matrizenaufgaben ab 2022 zur Studierendenauswahl in der Medizin eingesetzt werden sollen und kommerzielle Testvorbereitungsdienste nur selektiv den Bewerbenden zur Verfügung stehen, die über die notwendigen finanziellen Ressourcen verfügen. Es konnte gezeigt werden, dass bereits eine minimale Intervention direkt vor der Testbearbeitung zu einem großen Effekt führt, was zwar nicht die Konstruktvalidität jedoch durchaus die Testfairness beeinträchtigt. Diese Befunde fügten sich in die bestehende Literatur ein (Loesche et al., 2015; Schneider et al., 2020) wodurch deutlich wird, dass ein kostenloser Zugang zu Vorbereitungsmaterialien für alle Bewerbenden notwendig ist und keine Gefahr der Validitätsminderung besteht.

In Studie III (Koch, Spinath, et al., 2021) wurde die Notwendigkeit von figuralen Matrizen für die Studierendenauswahl erneut aufgegriffen und es wurde eine Itembank mit 219 IRT konformen figuralen Matrizen generiert. Diese Itembank sowie die dazu entwickelt Software stehen der Forschung kostenfrei zur Verfügung und können genutzt werden, um jährlich neue, aber gleich schwere Testversionen zu generieren, die einen dauerhaft fairen Auswahlprozess unterstützen. Das Itemmaterial basiert auf den gleichen Konstruktionsprinzipien wie die figuralen Matrizen in Studie II (Levacher et al., 2021), sodass angenommen werden kann, dass die gleichen Befunde zur Trainierbarkeit und Testfairness gelten.

Insgesamt konnte die vorliegende Dissertation viele positive Befunde der Forschung zur Intelligenzdiagnostik sammeln und teilweise mit neuen Daten unterstützen. Gleichzeit hat sie jedoch auch neue Probleme identifiziert und Lösungsvorschläge generierte und teilweise auch realisiert, damit auch in Zukunft die Bedeutung und Qualität von Intelligenzdiagnostik auf einem konstant hohen Niveau bleibt und gegebenenfalls sogar gesteigert werden kann. Hierfür sollte in Zukunft vor allem ein Fokus

auf interdisziplinäre Zusammenarbeit sowie die gemeinschaftliche Nutzung von Forschungswerkzeugen und -daten gelegt werden (z.B. Itembanken). Die vorliegende Dissertation hat dafür eine Grundlage geschaffen und verschiedene Anschlusspunkte aufgezeigt.

Literaturverzeichnis

Abad, F. J., Quiroga, M. A., & Colom, R. (2017). Intelligence Assessment. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.05599-1

Allison, B. Z., Wolpaw, E. W., & Wolpaw, J. R. (2007). Brain–computer interface systems: Progress and prospects. *Expert Review of Medical Devices*, *4*(4), 463–474. https://doi.org/10.1586/17434440.4.4.463

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Battauz, M. (2017). Multiple Equating of Separate IRT Calibrations. *Psychometrika*, *82*(3), 610–636. https://doi.org/10.1007/s11336-016-9517-x

Battauz, M. (2021). *equateMultiple: Equating of Multiple Forms*. https://cran.r-project.org/web/packages/equateMultiple

Becker, N., Schmitz, F., Falk, A., Feldbrügge, J., Recktenwald, D., Wilhelm, O., Preckel, F., & Spinath, F. (2016). Preventing Response Elimination Strategies Improves the Convergent Validity of Figural Matrices. *Journal of Intelligence*, *4*(4), 2. https://doi.org/10.3390/jintelligence4010002

Becker, N., & Spinath, F. M. (2014). *Design a Matrix Test. Ein Distraktorfreier Matrizentest zur Erfassung der Allgemeinen Intelligenz (DESIGMA)* [Measurement instrument]. Hogrefe.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of Serious Games: An Overview. *Advances in Human-Computer Interaction*, *2013*, 1–11. https://doi.org/10.1155/2013/136864

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research*, *16*(S1), 95–108. https://doi.org/10.1007/s11136-007-9168-6

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow Education, American Style: Test Preparation, the SAT and College Enrollment. *Social Forces*, *89*(2), 435–461. https://doi.org/10.1353/sof.2010.0105

Bundesverfassungsgericht, 1 Senat. (2017, December 19). *Bundesverfassungsgericht—Entscheidungen—Bundes- und landesgesetzliche Vorschriften über die Studienplatzvergabe für das Fach Humanmedizin teilweise mit dem Grundgesetz unvereinbar* (de) [Gerichtsentscheidung]. Bundesverfassungsgericht. https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2017/12/ls20171219_1bvl000314.html

Cadet, L. B., & Chainay, H. (2021). How preadolescents and adults remember and experience virtual reality: The role of avatar incarnation, emotion, and sense of presence. *International Journal of Child-Computer Interaction*, *29*. https://doi.org/10.1016/j.ijcci.2021.100299

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chituc, C.-M., Herrmann, M., Schiffner, D., & Rittberger, M. (2019). Towards the Design and Deployment of an Item Bank: An Analysis of the Requirements Elicited. In M. A. Herzog, Z. Kubincová, P. Han, & M. Temperini (Eds.), *Advances in Web-Based Learning – ICWL 2019* (Vol. 11841, pp. 155–162). Springer International Publishing. https://doi.org/10.1007/978-3-030-35758-0_15

Deutsch, J. E., & Westcott McCoy, S. (2017). Virtual Reality and Serious Games in Neurorehabilitation of Children and Adults: Prevention, Plasticity, and Participation. *Pediatric Physical Therapy*, *29*, S23–S36. https://doi.org/10.1097/PEP.0000000000000387

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Dodonov, Y. S., & Dodonova, Y. A. (2012). Response time analysis in cognitive tasks with increasing difficulty. *Intelligence*, *40*(5), 379–394. https://doi.org/10.1016/j.intell.2012.07.002

Dryburgh, E., McKenna, S., & Rekik, I. (2020). Predicting full-scale and verbal intelligence scores from functional Connectomic data in individuals with autism Spectrum disorder. *Brain Imaging and Behavior*, *14*(5), 1769–1778. https://doi.org/10.1007/s11682-019-00111-w

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Application of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation*, *9*(2), 158–161. https://doi.org/10.3758/BF03214491

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. Routledge, Taylor & Francis Group.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer-Based Tests on Racial-Ethnic and Gender Groups. *Journal of Educational Measurement*, *39*(2), 133–147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence*, *52*, 71–79. https://doi.org/10.1016/j.intell.2015.07.006

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3

Gottfredson, L. S. (2004). Intelligence: Is It the Epidemiologists' Elusive "Fundamental Cause" of Social Class Inequalities in Health? *Journal of Personality and Social Psychology*, *86*(1), 174–199. https://doi.org/10.1037/0022-3514.86.1.174

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Guo, M., Gutiérrez-Sánchez, P., Ortega-Álvarez, A., González-Calero, P. A., Quiroga, M. A., & Gómez-Martín, P. P. (2020). Telemetry and machine learning to speed-up the measure of intelligence through video games. *CoSECivi*, 160–171.

Haier, R. J. (1990). The end of intelligence research. *Intelligence*, *14*(4), 371–374. https://doi.org/10.1016/S0160-2896(05)80010-8

Haier, R. J. (2014). Increased intelligence is a myth (so far). *Frontiers in Systems Neuroscience*, *8*, 34. https://doi.org/10.3389/fnsys.2014.00034

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. *2014 47th Hawaii International Conference on System Sciences*, 3025–3034.

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*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14. https://doi.org/10.1016/j.intell.2014.10.005

Hearne, L. J., Mattingley, J. B., & Cocchi, L. (2016). Functional brain networks related to individual differences in human intelligence at rest. *Scientific Reports*, 8. https://doi.org/10.1038/srep32328

Hochschulrektorenkonferenz. (2021). Statistische Daten zu Studienangeboten an Hochschulen in Deutschland. Studiengänge, Studierende, Absolventinnen und Absolventen. Wintersemester 2021/2022. *Statistiken Zur Hochschulpolitik*.

https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-10-Publikationsdatenbank/Stat-2021-01_WS_2021_22.pdf

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hunt, E., & Pellegrino, J. (1985). Using interactive computing to expand intelligence testing: A critique and prospectus. *Intelligence*, *9*(3), 207–236. https://doi.org/10.1016/0160-2896(85)90025-X

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences*, 135–187. https://doi.org/10.1017/S0140525X07001185

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Koch, M., Becker, N., Spinath, F. M., & Greiff, S. (2021). Assessing intelligence without intelligence tests. Future perspectives. *Intelligence*, *89*. https://doi.org/10.1016/j.intell.2021.101596

Koch, M., Spinath, F. M., Greiff, S., & Becker, N. (2021). *Development and Validation of the Open Matrices Item Bank* [Manuscript submitted for publication].

Kokkinakis, A. V., Cowling, P. I., Drachen, A., & Wade, A. R. (2017). Exploring the relationship between video game expertise and fluid intelligence. *PLOS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0186621

Krautter, K., Lehmann, J., Kleinort, E., Koch, M., Spinath, F. M., & Becker, Nicolas. (2021). Test Preparation in Figural Matrices Tests: Focus on the Difficult Rules. *Frontiers in Psychology*, *12*, 7. https://doi.org/10.3389/fpsyg.2021.619440

Krieger, F., Zimmer, H. D., Greiff, S., Spinath, F. M., & Becker, N. (2019). Why are difficult figural matrices hard to solve? The role of selective encoding and working memory capacity. *Intelligence*, *72*, 35–48. https://doi.org/10.1016/j.intell.2018.11.007

Kroner, S., Plass, J., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*(4), 347–368. https://doi.org/10.1016/j.intell.2005.03.002

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*(2), 179–188. https://doi.org/10.1037/0033-2909.95.2.179

Lavery, M. R., Acharya, P., Sivo, S. A., & Xu, L. (2019). Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics - Simulation and Computation*, *48*(1), 27–38. https://doi.org/10.1080/03610918.2017.1371750

Levacher, J., Koch, M., Hissbach, J., Spinath, F. M., & Becker, N. (2021). You Can Play the Game Without Knowing the Rules – But You're Better Off Knowing Them: The Influence of Rule Knowledge on Figural Matrices Tests. *European*

Journal of Psychological Assessment, 1–9. https://doi.org/10.1027/1015-5759/a000637

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence*, *48*, 58–75. https://doi.org/10.1016/j.intell.2014.10.004

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. In *Handbook of Statistics* (Vol. 26, pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Primi, R. (2014). Developing a fluid intelligence scale through a combination of Rasch modeling and cognitive psychology. *Psychological Assessment*, *26*(3), 774–788. https://doi.org/10.1037/a0036712

Quiroga, M. A., Diaz, A., Román, F. J., Privado, J., & Colom, R. (2019). Intelligence and video games: Beyond "brain-games." *Intelligence*, *75*, 85–94. https://doi.org/10.1016/j.intell.2019.05.001

Quiroga, M. A., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., Gallego, B., & Colom, R. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can! *Intelligence*, *53*, 1–7. https://doi.org/10.1016/j.intell.2015.08.004

Quiroga, M. A., Román, F. J., De La Fuente, J., Privado, J., & Colom, R. (2017). The Measurement of Intelligence in the XXI Century using Video Games. *The Spanish Journal of Psychology*, *19*. https://doi.org/10.1017/sjp.2016.84

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. http://arxiv.org/abs/1711.05225

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Sachar, J., & Fletcher, J. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. *Proceedings of the 1977 Computerized Adaptive Testing Conference*, 403–419.

Santarnecchi, E., Emmendorfer, A., Tadayon, S., Rossi, S., Rossi, A., & Pascual-Leone, A. (2017). Network connectivity correlates of variability in fluid

intelligence performance. *Intelligence*, *65*, 35–47. https://doi.org/10.1016/j.intell.2017.10.002

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schmidt, F. L., & Hunter, J. (2016). *General mental ability in the world of work: Occupational attainment and job performance.* (p. 58). Sage Publications, Inc.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schneider, B., Becker, N., Krieger, F., Spinath, F. M., & Sparfeldt, J. R. (2020). Teaching the underlying rules of figural matrices in a short video increases test scores. *Intelligence*, 13.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Segovia, K. Y., & Bailenson, J. N. (2009). Virtually True: Children's Acquisition of False Memories in Virtual Reality. *Media Psychology*, *12*(4), 371–393. https://doi.org/10.1080/15213260903287267

Settanni, M., Azucar, D., & Marengo, D. (2018). Predicting Individual Characteristics from Digital Traces on Social Media: A Meta-Analysis. *Cyberpsychology,*

Behavior, and Social Networking, *21*(4), 217–228. https://doi.org/10.1089/cyber.2017.0384

Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2021). Good gamers, good managers? A proof-of-concept study with Sid Meier's Civilization. *Review of Managerial Science*, *15*(4), 957–990. https://doi.org/10.1007/s11846-020-00378-0

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality Prediction System from Facebook Users. *Procedia Computer Science*, *116*, 604–611. https://doi.org/10.1016/j.procs.2017.10.016

Tippins, N. T. (2015). Technology and Assessment in Selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

van der Linden, W. J., & Pashley, P. J. (2009). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3–30). Springer New York. https://doi.org/10.1007/978-0-387-85461-8_1

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement*, 20. https://doi.org/10.1177/0013164406288166

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments:

A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement*, *68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Ward, A. W., & Murray-Ward, M. (1994). An NCME Instructional Module: Guidelines for the Development of Item Banks. *Educational Measurement: Issues and Practice*, *13*(1), 34–39. https://doi.org/10.1111/j.1745-3992.1994.tb00785.x

Weiss, D. J. (2013). Item banking, test development, and test delivery. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* (pp. 185–200). American Psychological Association. https://doi.org/10.1037/14047-010

Wetzel, E., & Böhnke, J. R. (2017). Differential Item Functioning. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1297-1

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Interpreting Deep Visual Representations via Network Dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2131–2145. https://doi.org/10.1109/TPAMI.2018.2858759

Anhang

Studie I

Koch, M., Becker, N., Spinath, F. M., & Greiff, S. (2021). Assessing intelligence without intelligence tests. Future perspectives. *Intelligence*, 89.

https://doi.org/10.1016/j.intell.2021.101596

Studie II

Levacher, J.*, Koch, M.*, Hissbach, J., Spinath, F. M., & Becker, N. (2021). You Can Play the Game Without Knowing the Rules – But You're Better Off Knowing Them: The Influence of Rule Knowledge on Figural Matrices Tests. *European Journal of Psychological Assessment*. (*joint first authorship)

https://doi.org/10.1027/1015-5759/a000637

Studie III

Koch, M., Becker, N., Greiff, S., & Spinath, F. M. (2021). *Development and Validation of the Open Matrices Item Bank.* Manuscript submitted for publication.

https://doi.org/10.3390/jintelligence10030041